Disentangling mutation and selection in human genetic variation: promises and pitfalls

Ipsita Agarwal

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

ABSTRACT

Disentangling mutation and selection in human genetic variation: promises and pitfalls

Ipsita Agarwal

A subset of germline mutations that arise de novo each generation are deleterious and may cause severe genetic diseases. Predicting where in the genome and how often we expect to see deleterious mutations requires an understanding both of the distribution of mutation rates and the distribution of fitness effects in the genome. Both aspects are addressed in turn in the two projects described in this thesis.

The distribution of mutations in the genome is poorly understood because germline mutations occur very rarely. In Chapter 1 of this work, we investigated the sources of mutations by using the spectrum of low-frequency variants in 13,860 human X chromosomes and autosomes as a proxy for the spectrum of germline de novo mutations. By comparing the mutation spectrum in multiple genomic compartments on the autosomes and between the X and autosomes that have unique biochemical and sex-specific properties, we ascribed specific mutation patterns to replication timing and recombination and identified differences in the types of mutations that accrue in males and females. Understanding mutational mechanisms provides a basis for modeling mutation rate variation in the genome, which is ultimately needed to

infer the fitness effects of mutations.

In Chapter 2, we used patterns of human genetic variation at methylated CpG sites, known to experience mutations at very high rates, to directly learn about the fitness effects of mutations at these sites. In whole exome sequences now available for 390,000 humans, 99% of putatively-neutral, synonymous CpG sites have experienced a C>T mutation; at current sample sizes, not seeing a C>T mutation at these sites indicates strong selection against that mutation. We leveraged the saturation of neutral C>T mutations and the similarity of mutation rates at methylated CpG sites across annotations to identify the subset of sites in a given functional annotation of interest that are likely to be under strong selection. One implication of this work is that for the vast majority of sites in the genome, there will be little information about strong selection even in samples that are many times larger than at present; the distribution of fitness effects at highly mutable CpG sites may then serve as an anchor for what to expect for other types of sites.

Through the two specific cases described, this work illustrates the potential of large contemporary repositories of human genetic variation to inform human genetics and evolution, as well as their limitations in the absence of suitable models of mutation, selection, and other aspects of the evolutionary process.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to my advisor Molly Przeworski for her generous mentorship and guidance throughout the years, and for setting an incredible standard for me to aspire to as a scientist. My time in graduate school was made immeasurably richer by the privilege of learning from her, by her kindness and true inclusiveness, her constant support and copious edits.

I would also like to thank my committee members, Guy Sella, Itsik Pe'er, Peter Andolfatto and Kelley Harris; Guy and Itsik in particular for their thoughtful feedback and advice on various aspects of this work. For the rich intellectual environment I was fortunate enough to find myself in, I am deeply indebted to the current and former members of Przeworski, Andolfatto, and Sella labs. In particular, I'd like to thank Arbel Harpak, Hakhamanesh Mostafavi, Priya Moorjani, Molly Schumer, and Guy Amster, for many inspiring conversations about science, and their wise counsel at various points; Laura Hayward for her friendship and many memorable walks, Ana Pinharanda for being a wonderful officemate, and Felix Wu for many enjoyable conversations in the lab.

Finally, I would like to thank my parents for their unrelenting support over the

years; my sister Anoushka, and my cousins Malvika and Mishthi for being sources of joy and inspiration. I am also exceedingly grateful to many mentors who encouraged me over the years, and to my community of friends and family across five time zones without whom everything would be infinitely less meaningful.

To my grandmothers Kusum and Dev Bala.

Introduction

The diversity of extant life forms has arisen from a continual influx of genetic novelty in the form of new mutations and the pruning of mutations by natural selection and random genetic drift over time. For over a century, mathematical models that describe the evolution of genetic variation as a function of mutation, selection, drift, and recombination have been foundational to our understanding of the evolutionary process. These models have provided a framework, for instance, to ask how mutations that are neutral with respect to fitness and mutations that affect the fitness of organisms in their particular environments are expected to arise and accumulate, and how these mutations contribute to variation within and between species [1, 2, 3, 4, 5, 6].

It has been of longstanding interest to use these models to learn about the underlying processes of evolution from observed inter-individual variation [7], in part because of their potential to yield useful biological insights. The ability to sequence DNA at scale and directly examine genetic variation within and across species has revolutionized our ability to characterize genetic variation, and to infer parameters of interest in many cases. For instance, patterns of genetic variation among indi-

viduals sampled at present have been extensively used to learn about phylogenetic relationships [8], to infer demographic histories of populations [9, 10, 11], and to detect signatures of selection and thereby identify functionally important genomic loci and sites of important adaptations [12, 13, 14, 15, 16, 17]. Inferring individual parameters from genetic variation is difficult because it requires being able to model all the other parameters correctly, however. For instance, learning about the strength of selection at a site in the genome requires knowledge of the genealogical history and the mutation rate at the site, or a comparison to a designated set of neutral sites which can be assumed to have the same mutation rates and genealogical histories.

In turn, estimating the mutation rate correctly is contingent on being able to account for the effects of selection and drift, both of which remove mutations from the population. Early estimates of the mutation rate relied on the mutation-selection balance at strongly selected loci where drift could be neglected and selection could be estimated separately: in particular, Haldane obtained an estimate of the mutation rate from the observed frequency of hemophilia in the population, reasoning that new mutations that cause hemophilia arise every generation at a rate that balances out the observed reduction in fecundity of individuals with hemophilia [18]. The gold-standard approach for a long time involved calculating mutation rates from neutral divergence between humans and another species for which an estimate of the split time was available from the fossil record, relying on the rate of mutation being equal to the rate of neutral substitutions [19, 20]. More recently, the ability to sequence the genomes of thousands of parent-offspring trios has made it possible to identify de novo mutations in a single generation[21, 22]. Since these de novo mutations have only

undergone selection and drift for one generation, they faithfully reflect the mutational process: only mutations that are embryonic lethal would be missing from these data. De novo mutation data therefore provide the most direct possible estimate of the human mutation rate at present: $\sim 1.2 \times 10^{-8}$ per bp per generation for a typical site in the genome [21, 22].

A complication, which has been appreciated for some time, is that the biochemical propensity of DNA to experience mutations, differs at different types of sites in the genome; CpG dinucleotides, for instance, have long been known to have a very high mutation rate [23, 20]. Indeed, variability in mutation rates extends far beyond just CpG sites, and contributes substantially to differences in levels of genetic variation at sites across the genome, with so-called "cryptic" variation even within the same broad sequence contexts [24, 25]. The variation in mutation rates along the genome makes inference of selection more challenging, since the two contribute to genetic variation in uncharacterized ways at different sites. For example, the assumption that one could isolate selection at non-synonymous sites by comparing them with synonymous sites because these have the same mutation rate on average may not be warranted, even conditioning on broadly similar base compositions in those groups. This problem cannot be solved by obtaining direct de novo mutation rate estimates at a fine scale in the genome because de novo mutations are extremely rare events: in the largest study to date with 3000 samples, there are only about 200,000 de novo mutations in total, over 3 billion potential sites in the genome [26]. This has made it urgent and necessary to more generally understand the sources of mutation, and to use those to build mutational models that better predict this variability.

One approach to get at mutational mechanisms is to consider the "mutation spectrum", i.e., the distribution of different types of mutations, since different biochemical processes that generate or repair DNA lesions often leave distinct mutational signatures in DNA. In recent years, the types of data we can leverage to investigate mutational mechanisms have increased dramatically: in addition to recombination rates, replication timing and gene expression levels for numerous cell types, new methods to assay protein-DNA interactions and epigenetic features of sequences have yielded impressive amounts of information on methylation status, chromatin state, binding affinities of various proteins, and the incidence of other epigenetic markers at sites across the genome [27, 28, 29, 26, 30, 31]. Moreover, millions of somatic mutations have been sequenced in samples of both normal tissues and tumors – examining how the spectrum of somatic mutations varies with biochemical features such as replication timing and epigenetic modifications [32, 33, 34, 35, 36, 37, 38, 39, 40, 41], or across tumors of very different etiologies [42, 43, 44, 45] has been very useful in identifying mutational signatures of distinct mutagens and repair pathways. The key to using such an approach to understand sources of mutation in the germline is the ability to use polymorphism data as a substitute for extremely sparse de novo mutation data. In large samples, most polymorphisms are rare and recent enough for effects of direct and indirect selection and biased gene conversion to be minimal; they should therefore approximate the spectrum of germline mutations reasonably well [46, 47, 48]. Given that whole genome sequences are available for thousands of individuals [13, 14], the distribution of millions of rare variants across the genome can be used to investigate associations with various biochemical exposures.

Taking this approach in the first part of this work, we contrasted diversity levels of different mutation types across regions of the genome that differ with regard to specific biochemical properties to identify mutation types that are enriched in association with those biochemical properties. In comparing mutation spectrum differences across two compartments, we accounted for mean differences in mutation rates and genealogical histories across compartments. We found that C>A mutations are strongly associated with late replicating regions of the genome, suggesting that replication timing contributes substantially to variation across the genome for this mutation class. As another example, we showed that C>G mutations arise not only from repair of double-strand break damage in oocytes as previously suggested [22], but also from the repair of a subset of meiotic double-strand breaks in male germ cells. Thus, there is an appreciable effect of double-strand breaks on the mutation spectrum, localized in particular areas of the genome. Finally, relying on the notion that the X chromosome is enriched for female germline-specific biochemical exposures, we contrasted diversity levels on the X and autosomes, accounting for average X-autosome differences in replication timing and other features, to find numerous subtle sex-specific influences on the germline mutation spectrum.

Beyond insights into specific mutational mechanisms for some types of mutations, our results highlight that rates of different types of mutations are expected to vary substantially with local distributions of known biochemical features, as well as because of differences in sex-specific rates of damage and repair in different regions. Because we compared relative enrichments of mutation types and not mutation rates, we only learn about patterns of mutation rate variation in a qualitative sense. Nonetheless

this approach provides a starting point for examining relationships between other potential mutagenic exposures and mutation rates and for identifying features that best predict mutation rate variation for different types of mutations.

As an illustration for how information on mutational mechanisms and sources of variation in mutation rates can be useful, one could consider methylated CpG sites in the genome, where C>T mutations are known to occur at a rate ($\sim 1.2 \times 10^{-7}$ per bp per generation from de novo mutations) that is an order of magnitude higher than any other type of mutation in the genome [21]. At these sites, a single known mechanism, i.e., spontaneous deamination of methyl-cytosine, is believed to be predominantly responsible for their uniquely high mutability [23], and whether or not the CpG is methylated in the germline is a powerful predictor of the rate of C>T mutations [49]. By conditioning on methylation status, it is possible to identify groups of CpG sites that are highly mutable, and have similar distributions of mutation rates. Across such groups of sites then, assuming that they also have similar genealogical histories, patterns of genetic variation could be directly informative about the strength of selection.

Given that exome sequences are now available for hundreds of thousands of humans, we are approaching samples with genealogical histories long enough that in the absence of natural selection, every site with mutation rate on the order of $\sim 10^{-7}$ per bp per generation will have experienced at least one mutation over that time depth. In Chapter 2, we show that in a sample of 390,000 individuals, 99% of putatively-neutral, synonymous CpG sites with high levels of germline methylation have experienced a C>T mutation. We show through simulations that at current sample sizes, not see-

ing a C>T mutation at these sites indicates strong selection against that mutation. We leverage this mutation saturation of putatively neutral sites and the similarity of mutation rates at methylated CpG sites across annotations to identify the subset of sites in a given annotation that are likely to be under strong selection. We estimate, for example, that ~27% of loss of function mutations and ~6% of missense mutations are likely to be highly deleterious. One implication of this work is that for the vast majority of sites in the genome, we will be far from saturation even in samples that are many times larger than samples at present. Another is that fitness effects at highly mutable CpG sites may be a useful proxy for what to expect for the rest of the exome. More generally, we highlight the pitfalls of assigning pathogenic status to variants of unknown significance based on whether or not those mutations are observed in reference repositories of human genetic variation in the absence of a model.

In summary, I present here two specific illustrations of what can be learned about mutation and selection from large contemporary datasets, and the potential of such approaches to inform human genetics and evolution, as well as the current and future challenges associated with them.

Chapter 1

Signatures of replication timing, recombination and sex in the spectrum of rare variants on the human X chromosome and autosomes

Published under: Agarwal, I., Przeworski, M. (2019). *Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes.* Proceedings of the National Academy of Sciences, 116(36), 17916-17924.

1.1 Abstract

The sources of human germline mutations are poorly understood. Part of the difficulty is that mutations occur very rarely, and so direct pedigree-based approaches remain limited in the numbers that they can examine. To address this problem, we consider the spectrum of low frequency variants in a dataset (gnomAD) of 13,860 human X chromosomes and autosomes. X-autosome differences are reflective of germline sex differences, and have been used extensively to learn about male versus female mutational processes; what is less appreciated is that they also reflect chromosome-level biochemical features that differ between the X and autosomes. We tease these components apart by comparing the mutation spectrum in multiple genomic compartments on the autosomes and between the X and autosomes. In so doing, we are able to ascribe specific mutation patterns to replication timing and recombination, and to identify differences in the types of mutations that accrue in males and females. In particular, we identify C>G as a mutagenic signature of male meiotic double strand breaks on the X, which may result from late repair. Our results show how biochemical processes of damage and repair in the germline interact with sex-specific life history traits to shape mutation patterns on both the X chromosome and autosomes.

1.2 Introduction

Germline mutations, the source of all heritable variation, accrue each generation from accidental changes to the genome during the development of gametes. These mutations reflect a balance of exogenous damage or endogenous processes that alter

DNA in the germline, and processes that correctly repair DNA lesions before the next replication [50]. The biochemical machinery that underlies germline mutagenesis can be conceptualized as a set of genetic loci that modulate the net mutational input in each generation, and variants in these loci as "modifiers of mutation" [51, 52]. Since the activity of distinct biochemical pathways often leaves different signatures in DNA [43, 42, 53, 54, 44, 45, 55], these modifiers influence the distribution of mutation types (the "mutation spectrum"), as well as the total number of mutations inherited by offspring.

The mutational landscape in the germline is also modified by the sex of the parent: in humans, notably, it has long been known that males contribute three times as many mutations on average as females per generation [56, 21]. As in other mammals, gametogenesis differs drastically by sex: female germ cells are arrested in meiosis for much of their development whereas male germ cells enter meiosis late in their development [57, 58, 59]. Male germ cells undergo many more cell divisions than female germ cells; they are also methylated earlier and have higher methylation levels on average throughout ontogenesis [60]. Due to differences in their cellular biochemistry at different developmental stages, male and female gametes may be subject to different kinds of endogenous and environmental insults, or repair different types of DNA lesions with varying degrees of efficacy. For example, male gametes may accrue oxidative damage due to lack of base excision repair in late spermatogenesis [61]. Males and females also differ in life history traits such as the timing of puberty and age of reproduction [62], which modulate the exposure of the gamete to the biochemical states associated with particular stages of development and thus alter their muta-

genic impact. In that sense, the sex of the parent as well as variants in loci associated with sex-specific biochemistry and life history are also modifiers of mutation. The germline mutation spectrum in each generation is therefore a convolution of the signatures left by biochemical machinery in DNA sequence and the influence of sex on the developmental trajectories of germ cells.

In principle, it is possible to characterize mutational mechanisms by decomposing the mutation spectrum into its component signatures. Such an approach has led to a wealth of insight into the sources of somatic mutations, i.e., mutations that accumulate in somatic tissues during normal development or ageing. Distinct signatures of processes that generate or repair DNA lesions have been identified by analyzing millions of somatic mutations in their immediate sequence context, across tumor samples of diverse etiologies [42, 43, 44, 45]. A complementary approach, based on changes in the mutation spectrum with regional variation in genomic features, has further illuminated the influence of local replication timing, transcription, chromatin organization, and epigenetic modifications on somatic mutagenesis [32, 33, 34, 35, 36, 37, 38, 39, 40, 41].

These methods have proved difficult to apply to the germline however, because each offspring inherits only about 70 de novo mutations on average [22]. Thus, the most direct approach to the study of germline mutations, the resequencing of pedigrees [63, 64, 22, 21, 47], remains limited in its ability to identify determinants of mutation rate variation. For instance, examining 96 possible mutation types considered in a trinucleotide context in ~100,000 de novo mutations, the biggest study to date found only three mutation types for which the proportion transmitted from

mothers and fathers differed significantly ([22]). Additionally, the mutation patterns from the three largest de novo mutation studies combined show inconsistent patterns of correlation to genomic features, for reasons that remain unclear [65].

One way to overcome the limitation of small samples in studies of germline mutation is to use polymorphisms as a proxy for de novo mutations. In large samples, most polymorphisms are rare and recent enough for effects of direct and indirect selection and biased gene conversion to be minimal; they should therefore recapitulate the de novo mutation spectrum with reasonable fidelity [46, 47, 48]. The much higher density of rare variants across the genome can then be used to more robustly investigate associations with genomic features. Using this strategy, a recent study of human autosomal data identified mutation types and contexts significantly associated with a variety of genomic features [46]. While the authors suggested putative biochemical sources for three signatures in the germline based on their similarity to patterns that have been reported in tumors, it is unclear to what degree these mechanisms can be directly extrapolated to the germline [66, 67]. Moreover, sex-specific effects on the mutation spectrum were not considered.

Insight into sex-specific effects can be gained by contrasting polymorphism levels on the sex chromosomes and autosomes, since autosomes reflect mutational processes in the male and female germlines equally, while the X chromosome disproportionately reflects the female germline, and the Y chromosome exclusively reflects the male germline. This approach to studying sex differences has been used extensively; notably, its application to divergence data provided the first systematic evidence for a higher contribution of males to mutation in humans and other mammals [68, 69].

Yet no significant influence of sex on the mutation spectrum was inferred in a recent comparison of ~3000 rare variants on the X and Y chromosomes [47]. Despite their importance, therefore, the genesis of germline mutations remains poorly understood to date, and the role of sex-specific modifiers particularly enigmatic.

To fill this gap, we consider the spectrum of rare polymorphisms across the genome using genome-wide SNPs in the gnomAD dataset [70, 14]. We compare particular genomic "compartments", or units of the genome with unique combinations of biochemical and sex-specific properties, on the X and autosomes; this approach enables us to tease apart biochemical and sex-specific influences on the germline mutation spectrum. With over 120 million SNPs to analyze across the genome, we can thus detect even subtle differences in mutational patterns between genomic compartments.

1.3 Materials and Methods

We use whole genome SNP data from 15,496 individuals made available by the Genome Aggregation Database (gnomAD), which includes 9,256 Europeans and 4,368 African or African-American individuals [70, 14]. We limit our analysis to the 6,930 female individuals in the dataset to sample X-chromosomes and autosomes in equal numbers. We then compare the diversity levels of different mutation types in pairs of genomic compartments (**Figure 1.1**). In these data, there are ~120 million SNPs, of which 53% of the variants are singletons (i.e., variants seen only once in the sample, with an allele count of one), and 11% are doubletons (allele count = 2) (**Figure 1.1**). Only about 10% of variants are at frequency 1% or greater; we retain them, given

that their inclusion does not affect our qualitative results (Supplementary Methods, **Figure 1.6**).

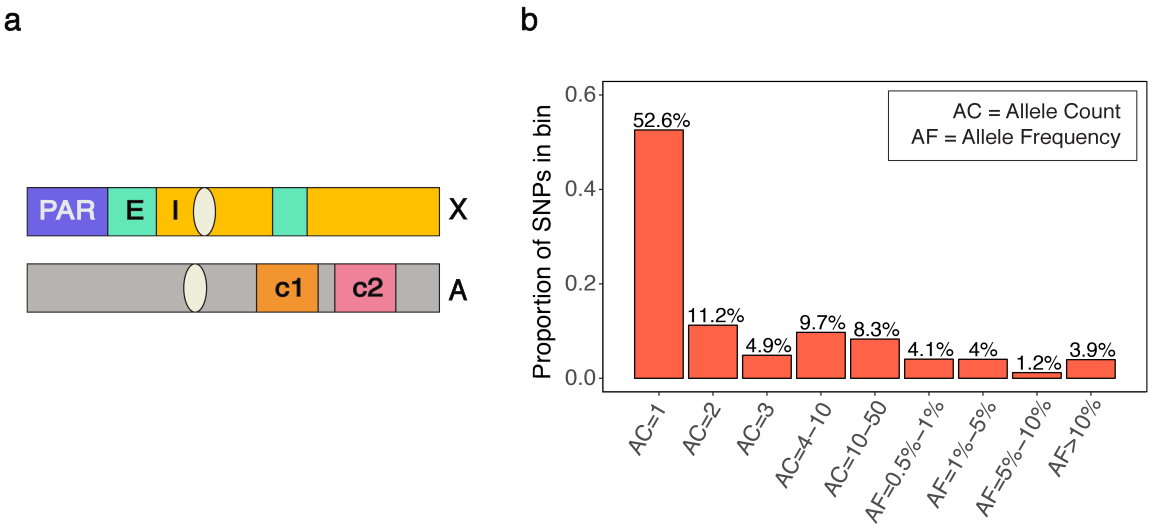

Figure 1.1: **(a)** A schematic of genomic compartments on the X chromosome and autosomes. Three compartments on the X chromosome are depicted: the pseudoautosomal region (PAR), regions of the X that escape inactivation (E), regions of the X that undergo inactivation (I). Also depicted are two hypothetical autosomal compartments with distinct biochemical properties (c1 and c2). Analyses include pairwise comparisons of mutational patterns between autosomal compartments and between the X chromosome and autosomes. **(b)** The frequency spectrum of variants (N = 120,521,915) in the 13,860 chromosomes analyzed. Over two-thirds of mutations are present at three copies or less in the sample.

As in other recent studies, we extract the single base pair flanking sequence on each side of the variant position using the hg19 reference to obtain mutations in their trinucleotide context, and combine mutations in reverse complement classes (for example, the ACG>ATG and CGT>CAT classes are collapsed into the former) to obtain 96 mutation types. Unless otherwise noted, we treat the major allele as the ancestral state at a site; however, we obtain similar results using the ancestral allele and context from the 1000G reconstruction of the ancestral human genome sequence [71] (Supplementary Methods, **Figure 1.7**). We include multi-allelic sites (~6% of the data) by counting the multiple derived alleles separately as if they had occurred

at separate bi-allelic sites with the same major allele (Supplementary Methods, **Figure 1.8**). To obtain the diversity level for each mutation type within a genomic compartment, we divide the number of segregating sites of a particular type by the number of mutational opportunities, i.e., sites where a single change could have given rise to that mutation type; this approach accounts for base composition within a compartment.

To compare mutation types across two genomic compartments, we normalize the diversity for each mutation type by the total diversity within each compartment. In this way, we control for the effect of population genetic processes that affect diversity across compartments but do so evenly across all mutation types, and isolate differences in the mutation spectrum; this step is particularly important for comparisons between the X chromosome and autosomes. For each of 96 mutation types, we test if the observed relative diversity in the two compartments differs from what would be expected by chance. To this end, we designate one of the two compartments as the "test" and the other as the "reference" compartment. Our null expectation is that the number of mutations of a particular type in the test compartment is binomially distributed with a mean value proportional to the observed diversity for that type in the reference compartment, adjusted for overall differences in diversity between the two compartments (Supplementary Methods). Mutation types are considered significantly different in their frequencies between the two compartments if the two-tailed p-value from the binomial test is below the Bonferroni-corrected 5% significance threshold. This approach implicitly ignores sampling error in the estimate of diversity of the designated reference compartment; we verify that our results are insensitive

to this assumption by using alternative approaches to calculate significance that do not make this assumption, but have other limitations (Supplementary Methods). We consider the effects of highly mutable types on the distribution of other mutation types (Supplementary Methods; **Figure 1.9**); we also consider possible differences in sequencing error rates between compartments, and replicate our findings in two alternate datasets (Supplementary Methods; **Figure 1.10-Figure 1.12**).

1.4 Results and Discussion

Biochemical properties vary along the genome, both on autosomes and the X chromosome. In turn, sex-specific influences from the germline are the same across autosomes, but differ between the X chromosome and autosomes. We therefore first compare autosomal compartments with distinct biochemical features to illuminate biochemical influences on the mutation spectrum. Then, by comparing compartments across the X chromosome and autosomes and accounting for average biochemical differences between them, we disentangle sex-specific and biochemical influences on the mutation spectrum.

1.4.1 Replication timing and its covariates influence the germline mutation spectrum

We consider autosomal compartments that differ with regard to specific biochemical properties in the germline. In cases where these data are unavailable for germline tissue and we are limited to somatic cell lines, we focus on biochemical features that

have broadly similar distributions across tissue types. Replication timing is consistently an important predictor of local mutation rates [72, 65, 66] in both the soma and the germline, and broad-scale replication timing maps are relatively concordant across tissues [27, 73] (**Figure 1.13**). The observed mutagenic effect of late replication has been hypothesized to be due to a decline in the efficacy of mismatch repair with delayed replication, less time for repair, or the accumulation of damage-prone single-stranded DNA at stalled replication forks [72, 40].

To assess if replication timing affects the germline mutation spectrum, we compare autosomal regions that differ in their replication timing using available data from LCL and H9-ESC cell lines [27, 28]. As expected, almost all mutation types are significantly enriched in late replicating regions relative to early replicating regions (**Figure 1.2**, Fig. S9a). In particular, we observe a substantial enrichment of C>A and T>A mutations in late replicating regions, a pattern also observed by Carlson et al., 2018 in a different sample of rare variants. Moreover, the mean replication timing in 1 Mb windows across the genome explains ~60% of the variation in C>A and T>A enrichment in those windows relative to the autosomal average and between 2% and 26% for all other mutation types ($p << 10^{-5}$), suggesting that these two mutation types are particularly sensitive to replication timing ((**Figure 1.6**, **Figure 1.14**b, SI Appendix).

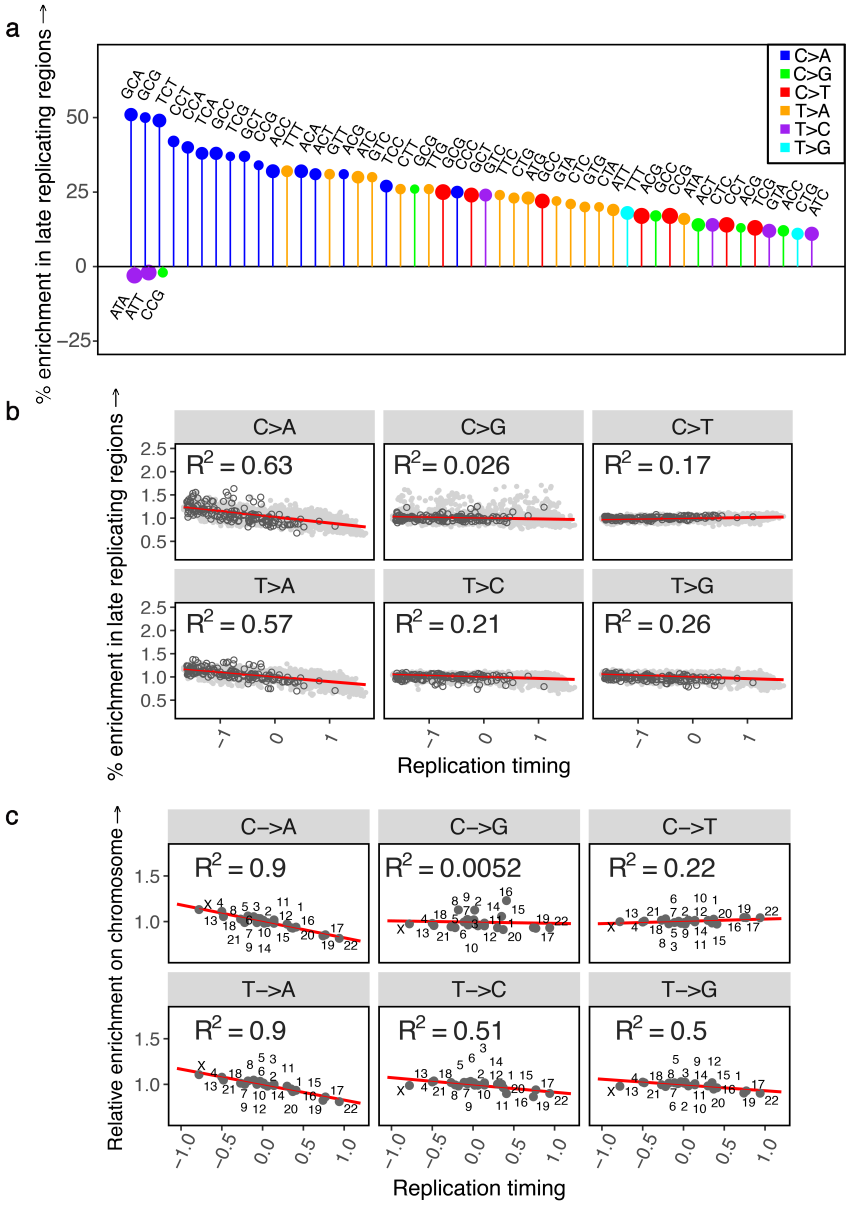

Figure 1.2: The effect of replication timing on the mutation spectrum at different scales, using replication timing scores from the LCL cell line. **(a)** Comparison of the spectrum of 96 mutation types in late replicating (score <= -0.5) autosomal regions relative to early replicating (score >= 0.5) autosomal regions. Positive and negative effects have been separately ordered by effect size from left to right; only the top 50 significant positive and negative effects are shown for legibility. The size of the circle reflects the number of mutations of that type. **(b)** For each of six mutational classes, the enrichment in 1Mb windows relative to all other autosomal windows combined, ordered by the mean replication timing. Positive replication timing scores indicate earlier than average replication. Autosomal windows are shown in solid light grey circles; windows on the X chromosome have been overlaid in black hollow circles. **(c)** For each of six mutational classes, enrichment on individual autosomes and X relative to all other autosomes combined, ordered by the mean replication timing. Positive replication timing scores indicate earlier than average replication.

Because replication timing is correlated with multiple genomic features, including higher order chromatin structure, epigenetic modifications, and in particular, DNA methylation at CpG sites, some of the observed patterns could be reflective of these processes rather than replication per se. To assess the marginal impact of CpG methylation on the effect of replication timing, we consider early and late replicating regions within and outside CpG islands, which are regions of CpG hypomethylation across tissue types [74, 75]. We find that at both CpG sites inside and outside islands, C>A mutations are enriched in late replicating regions (**Figure 1.15**), suggesting that this signal is not due to differences in methylation. Moreover, we also observe this pattern at non-CpG sites (**Figure 1.15**).

The association of C>A and T>A mutations with replication timing does not necessarily imply that they are "replicative" in origin, i.e., due to errors directly introduced by the replication machinery while copying intact DNA, as they could also reflect greater unrepaired damage in later replicating regions [40]. In particular, since C>A mutations are a known consequence of oxidative damage in somatic tissues [42, 76, 77, 78, 43], it is plausible that these mutations accumulate in regions of late replication due to greater damage to exposed single-stranded DNA, or poorer repair in these regions.

Considering other factors shown to influence mutation patterns, we recover a known signature of CpG methylation: transitions at CpG sites (C>T mutations in the ACG, CCG, GCG, and TCG trinucleotide contexts), which are thought to be due to the spontaneous deamination of methyl-cytosine to thymidine, are highly depleted in the hypomethylated CpG islands compared to the rest of the genome (**Figure**

1.16). Similarly, we detect an increase in C>G mutations in a subset of autosomal regions previously shown to be enriched for clustered C>G de novo mutations (**Figure 1.16**b). This C>G signature is thought to reflect inaccurate repair of spontaneous damage-induced double-strand breaks in the germline [22, 49].

Importantly, the impact of these biochemical features on mutation does not average out across chromosomes. Comparing individual autosomes to all other autosomes reveals ubiquitous variation in the mutation spectrum at the chromosome-level (**Figure 1.16**). In particular, individual chromosomes that replicate later on average show greater enrichment of C>A and T>A mutation types: differences in mean replication timing for individual autosomes explain ~90% of the variation in C>A and T>A enrichment at the chromosome level ($p << 10^{-5}$), while they explain ~50% or less for other mutation types (**Figure 1.2**). These results demonstrate that replication timing, and potentially other genomic features such as methylation and propensity for accidental double strand break damage, lead to chromosome-level differences in diversity, hinting at some plausible sources for observed but unexplained chromosome-level differences in average divergence [67].

1.4.2 Sex-specific influences on the mutation spectrum are subtle but likely ubiquitous

Next, we assess the impact of sex on the germline mutation spectrum by comparing mutational patterns on the X chromosome and autosomes. The X chromosome is disproportionately exposed to mutational processes in the female germline; viewed

from a population perspective, there are more X chromosomes in females than in males, but the same number of autosomes in both. Thus, mutation types that arise more commonly in the female germline are expected to be enriched (and mutation types that arise more commonly in the male germline depleted) on the X chromosome relative to autosomes. We account for population-level properties that may affect the mutation spectrum differently on the X and autosomes (Supplementary Methods). Having done so, we find most mutation types to be differentially enriched on the X and autosomes (**Figure 1.3**).

Importantly, however, these X-autosome differences do not only reflect differences in male and female mutational processes; given the substantial effect of biochemical features on mutational patterns observed at the chromosomal scale, they also potentially reflect differences in the distribution of these biochemical features on the X chromosome and autosomes. For instance, in de novo mutation studies [64, 22], C>A mutations are found to arise more often in males, suggesting that they should be depleted on the X. Instead, they are found enriched on the X chromosome relative to autosomes (**Figure 1.3**). A possible explanation is that the X accrues excess C>A mutations because it replicates late in the germline. C>A mutations are known to be associated with oxidative damage [42, 76, 77, 78, 43], which remains unrepaired in sperm [61], and is likely repaired at or before the first cell division in the zygote [79, 80, 81]. Late replication of the X chromosome at this stage, perhaps due to the inactive status of the paternally inherited X in female embryos [82], could then indeed be expected to result in an enrichment of C>A mutations on the X relative to autosomes, despite a primarily male source of damage. This example underscores

that accounting for the X-specific effects of biochemical features is key to uncovering true sex differences in X-autosome comparisons.

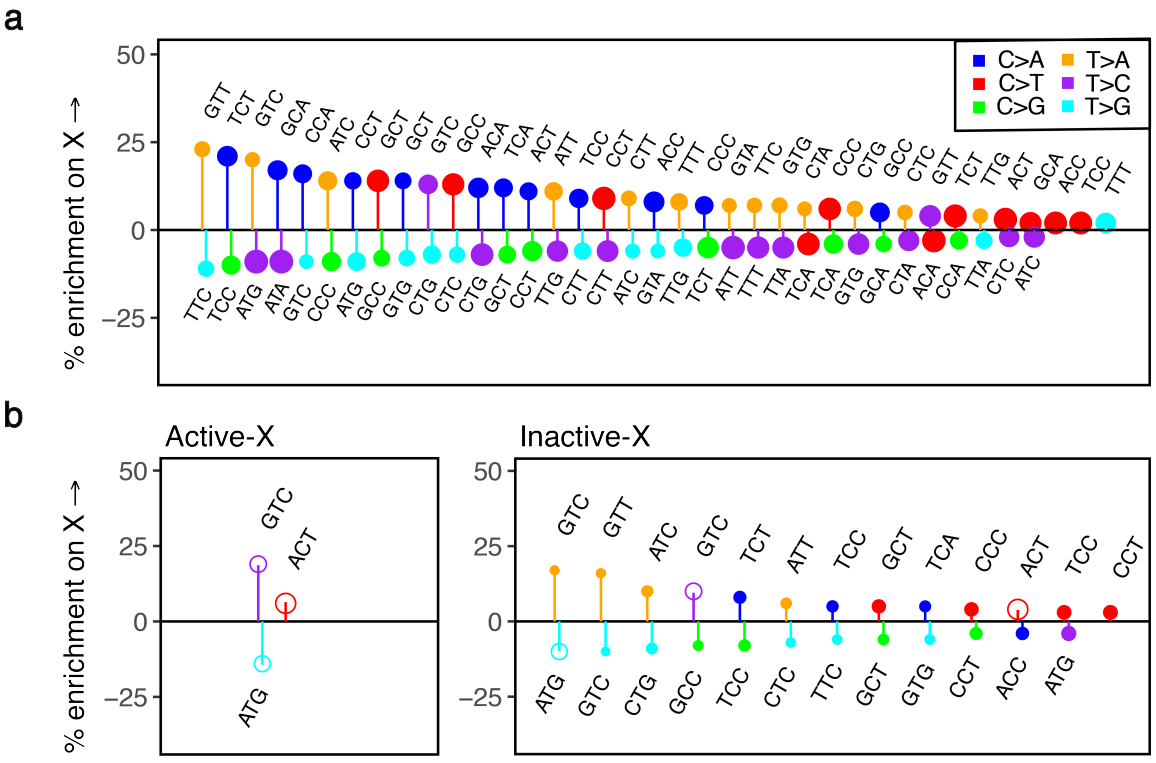

Figure 1.3: Comparison of the mutation spectrum on the X chromosome and autosomes. The pseudoautosomal region (PAR) and CpG sites are excluded from this analysis. Only significant differences are shown. Positive and negative effects have been separately ordered by effect size from left to right. The size of the circle reflects the number of mutations of that type. **(a)** Enrichment of mutation types on the X chromosome relative to autosomes. **(b)** Enrichment of mutation types in the genic compartment of the X chromosome that escapes inactivation, relative to genic regions on autosomes. Hollow circles represent mutation types enriched (or depleted) in both the escaped (active) and inactive compartments of the X relative to autosomes. **(c)** Enrichment of mutation types in the genic X chromosome compartment that undergoes X-inactivation, relative to genic regions on autosomes. Hollow circles represent mutation types enriched (or depleted) in both the escaped (active) and inactive compartments of the X relative to autosomes. The larger number of significant differences in (c) compared to (b) likely reflects at least in part the approximately five-fold greater amount of data in the inactive versus the active genic regions of the X chromosome.

One well-characterized idiosyncratic property of the X is X-inactivation, which is associated with X-specific changes in methylation, transcriptional activity, and notably, replication timing: because the inactive X chromosome exhibits a significant

lag in replication, on average the X replicates later than autosomes [83]. Though X-inactivation is a short-lived process in the germline—limited to early embryogenesis in females, and brief meiotic and post-meiotic periods in males [84, 85, 86, 87]—it could nevertheless lead to observable differences in the mutation spectrum between different regions of the X. The "active" compartment of the X chromosome, i.e., the approximately 15% of the transcribed X that constitutively escapes inactivation across tissues [88, 89] may therefore differ in its mutation spectrum from the rest of the X. Comparing autosomes with the inactive and active regions of the X, we find T>C mutations at GTC sites and C>T types at ACT sites enriched in both active and inactive regions of the X relative to autosomes and T>G mutations at ATG sites depleted both in the active and inactive regions of the X relative to autosomes (**Figure 1.3**, **Figure 1.17**). Since these cases cannot be attributed to X-inactivation and are enriched (or depleted) concordantly on compartments of the X chromosome that differ in their replication timing, methylation levels and other features, they are strong candidates for true sex differences in mutation. Given that the genic compartment known to escape inactivation across tissues is a small fraction of the X chromosome, there are likely many more subtle ones that we miss.

A complementary approach to minimizing X-specific biochemical influences on the mutation spectrum of the X in X-autosome comparisons is to consider regions of the X chromosome that are comparable to autosomes in their average replication timing. The replication timing on the X chromosome across multiple human cell lines depends on whether one of the X chromosomes is inactivated (**Figure 1.13**; Supplementary Methods) [90, 91, 92, 73, 27, 59]. This observation suggests that

controlling for replication timing differences between the X chromosome and autosomes may also control for the effects of other correlated features, including those associated with X-inactivation. Using this approach, all three mutation types that we highlight as putative sex differences based on their differential enrichment in the active compartment of the X relative to autosomes are also observed as significant differences between the X chromosome and autosomes (**Figure 1.4**, **Figure 1.18**). That we find the same types with this complementary approach provides further evidence that they are true sex differences.

We also detect a number of additional differentially enriched types between X and autosomes after controlling for replication timing differences (**Figure 1.4**); many of these types are enriched concordantly in early and late replicating regions of the X relative to autosomes (**Figure 1.19**). Assuming that a majority of X-specific effects are accounted for when we control for replication timing, these types can also be considered putative sex differences. In that respect, it is noteworthy that C>A mutations are enriched in inactive or late replicating regions of the X, but depleted in the active or early replicating regions of the X, when compared to autosomes (**Figure 1.17**,**Figure 1.19**). This pattern is what we would expect from the combined influences of a male bias and an effect of replication timing on C>A mutations, as we suggested earlier.

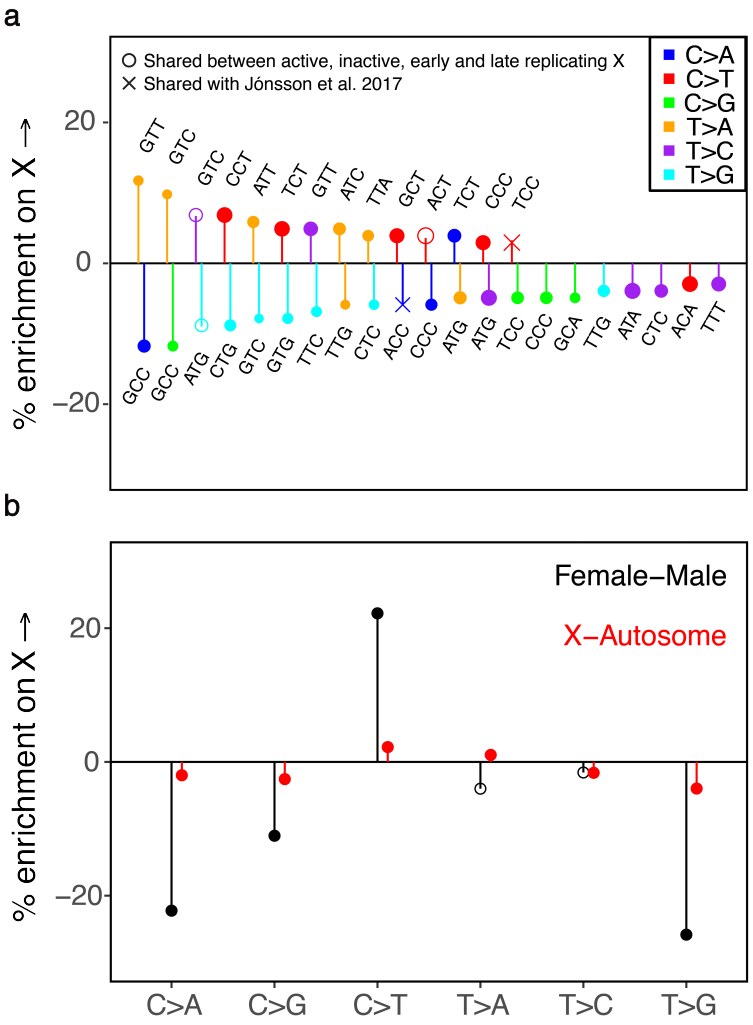

Figure 1.4: The mutation spectrum on the X and autosomes matched for average replication timing. The pseudoautosomal region (PAR) and CpG sites are excluded from this analysis. **(a)** Comparison of the mutation spectrum on the X chromosome and autosomes matched for average replication timing. Only significant differences are shown. Positive and negative effects have been separately ordered by effect size from left to right. Hollow circles represent mutation types also enriched (or depleted) in both the active and inactive compartments of the X relative to autosomes. Crosses denote mutation types reported as significant sex differences by Jonsson et al., 2017. **(b)** The X-autosome spectrum for six mutation classes, controlling for mean replication timing (in red), compared to known male female differences from Jonsson et al., 2017 (in black). Solid points are statistically significant differences at the 5% level, accounting for multiple tests.

We further assess these putative sex-specific signatures by comparing them to results from the largest human pedigree study of de novo mutations to date ([22]). Among the six broad mutational classes, Jonsson et al. find C>T mutations sig-

nificantly enriched in maternal, and C>A, C>G, and T>G mutations relatively enriched in paternal de novo mutations. The mutational patterns we observe on the X chromosome and autosomes after controlling for differences in replication timing are consistent with these effects: we find C>T mutations significantly enriched and C>A, C>G, and T>G classes significantly depleted on the X chromosome relative to autosomes ((**Figure 1.4**, **Figure 1.18**). Jonsson et al. also find three mutation types in their trinucleotide context (TCC>TTC, ACC>AAC, ATT>AGT) as significant sex differences: of these we find two as significant X-autosome differences. As expected, the maternally enriched TCC>TTC type is relatively enriched on the X chromosome, and the paternally enriched ACC>AAC type is relatively enriched on autosomes (**Figure 1.4**). We do not observe the third type as differentially enriched on the X and autosomes, possibly because there are genomic features specific to the X that mask its enrichment in females.

In turn, the types that we identify as putative sex differences from the comparison of X active, X inactive and autosomes are not reported as significant sex differences in Jonsson et al. (2017). The reason may be that most of them reflect subtle X-autosome differences, with X-enrichment or depletion in the range of 5-10%. Translating these enrichments into a difference between males and females requires a full population genetic model, including assumptions about demography and life history [93]. Nonetheless, such subtle X-autosome differences likely correspond to small sex differences that current de novo studies are underpowered to detect.

1.4.3 A subset of meiotic double-strand breaks have the same mutagenic impact as accidental damage

In the preceding section, we suggested a plausible mechanism through which local biochemical influences and sex-specific properties of the germline jointly influence the distribution of C>A mutations on the X chromosome and autosomes. Here we highlight another mutation type, C>G, which is also distributed in a sex-specific and chromosome-specific manner, but is largely insensitive to replication timing.

As recently reported, clustered C>G de novo mutations are concentrated in particular autosomal regions, and the number of such mutations transmitted in each generation increases exponentially with maternal age at reproduction [22, 49]. Maternal age at reproduction determines the duration of oocyte arrest, since females are born with their entire complement of oocytes, which remain in dictyate arrest until ovulation. Based on the sex-specific patterns of accumulation with age and genomic properties of these mutations, the authors speculated that the C>G clusters could be due to the more frequent spontaneous occurrence of damage-induced double strand breaks (DSBs) in some genomic regions and an increasing rate of such damage in older oocytes. In this view, C>G mutations are associated with accidental double-strand break damage in both males and females.

Accidental damage is not the only source of double strand breaks in the germline, however: during meiosis, double strand breaks are deliberately induced along the genome, through targeting of PRDM9-binding motifs [94, 95]. These DSBs are repaired through the homologous recombination pathway: a small minority are resolved

through crossovers (COs), which involve exchanges of large segments between homologous chromosomes, and the rest are thought to be repaired through non-crossover gene conversion events (NCOGCs), though another small subset may involve non-homologous end joining and other mechanisms [96, 97]. Potentially, these meiotic DSBs could have a mutagenic impact similar to that of spontaneous double-strand breaks; however, a clear mutational pattern common to both has not been seen to date. For instance, using DMC1 ChIP-Seq data from human spermatocytes, Pratto et al., 2014 observed C>G enrichment to a small degree around male autosomal hotspots [29], but the source of these types was not discussed further by the authors, and is potentially due to overlap with regions of clustered de novo C>G mutations reported by Jonsson et al., 2017. Another recent study did not find de novo C>G mutations enriched within autosomal crossover hotspots identified in pedigree studies [26]. We test if there is indeed an enrichment of C>G mutations associated with meiotic DSBs by comparing the mutation spectrum within and outside hotspots on autosomes; we use DMC1 hotspots in males and crossover hotspots in females because we do not have a map of DMC1-binding in female gametes. Our results are consistent with previous observations: we do not observe C>G enrichment in autosomal hotspots for males or females once we exclude regions of clustered de novo C>G mutations (**Figure 1.16**, **Figure 1.20**).

We next consider the X chromosome, which in females recombines like an autosome, but in males is in the unusual position of having no homolog outside the pseudoautosomal region (PAR). In males as in females, meiotic DSBs are nonetheless made both inside and outside the PAR [98, 99, 100, 29] Properties of recombination

events on the X and autosomes differ markedly between sexes however: notably, the pseudoautosomal region 1 (PAR1), a 2.6 Mb region on the X chromosome, experiences an obligate crossover in males, but normal levels of recombination in females [101, 98, 102, 103] and in male germ cells, DSBs are repaired late on the X chromosome relative to autosomes [98, 102, 99, 100, 29]. These considerations raise the possibility that mutational patterns in hotspots on the X chromosome in males may reflect these sex-specific features of recombination and behave differently relative to autosomal hotspots in males, and relative to both X and autosomes in females. To explore this hypothesis, we compare mutation patterns on autosomes to those in PAR1, which is exposed to the male and female germlines to the same degree as autosomes (since two copies are carried by both males and females), and does not undergo X-inactivation [104]. We find that C>G mutation types are systematically enriched on the PAR1 relative to autosomes (**Figure 1.5**), indicating that repair of meiotic double-strand breaks in this region in males is associated with C>G enrichment.

We further characterize the source of the C>G enrichment using DMC1 ChIP-Seq data from human spermatocytes [29]. The DMC1 signal reflects intermediates in the homologous recombination pathway; high levels of DMC1-binding can reflect either an increased frequency of double strand breaks (hotspots of greater intensity) or a greater duration of intermediates, i.e., a longer time to repair [98, 29]. Using these data, we find that there is clear C>G enrichment not only in PAR1, but also in hotspots on the X chromosome outside PAR1; moreover, the enrichment increases with the strength of the DMC1 signal ((**Figure 1.5**, **Figure 1.20**).

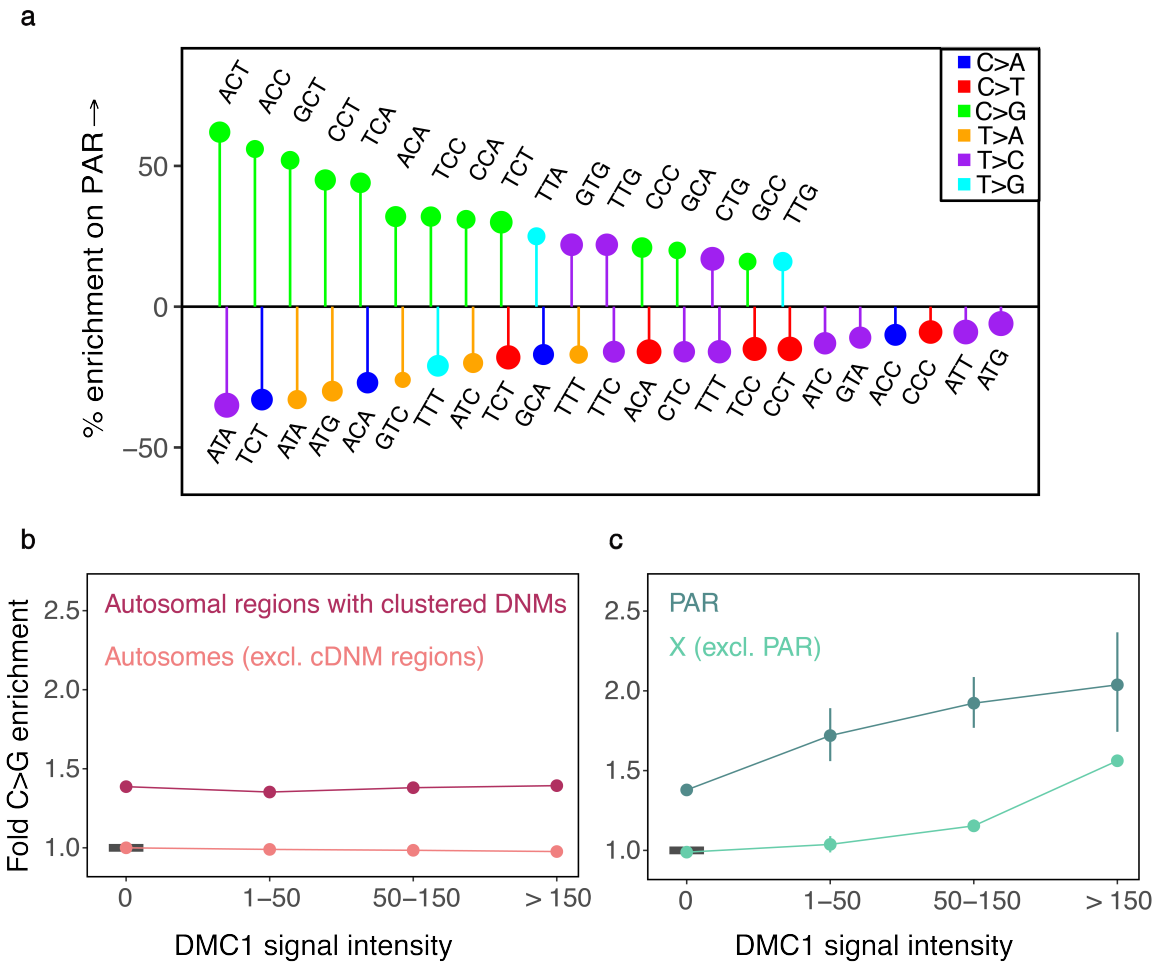

Figure 1.5: Distribution of C>G mutations in genomic compartments relative to autosomes. CpG sites are excluded from these analyses. **(a)** The mutation spectrum on PAR1 relative to autosomes. Only significant differences are shown. Positive and negative effects have been separately ordered by effect size from left to right. **(b)** Enrichment of C>G mutations in DMC1 hotspots of varying intensity on autosomes. For each estimate, the 95% confidence interval from a binomial test is represented by the vertical bars (and is sometimes too small to be apparent). The horizontal black bar shows the reference, namely autosomes outside DMC1 hotspots and excluding regions rich in clustered de novo mutations (identified by Jonsson et al., 2017). **(c)** Enrichment of C>G mutations in DMC1 hotspots of varying intensity on the X chromosome, relative to autosomes outside DMC1 hotspots and excluding autosomal regions rich in clustered de novo mutations. For each estimate, the 95% confidence interval from a binomial test is represented by the vertical bars. The reference, same as for (b), is denoted by a horizontal black bar.

That we observe C>G enrichment in male hotspots on the X chromosome but not in male hotspots of similar average intensities on autosomes (**Figure 1.5**, **Figure 1.20**) or in female crossover hotspots on autosomes (**Figure 1.20**), leads us to specu-

late that the predominant source of this C>G signature is the delay in repair of DSBs on the X chromosome relative to autosomes in male meiosis. We note that because hotspots detected in spermatocytes could also be used in female meiosis [105], without a female-specific map of DMC1 binding, we cannot exclude the possibility that C>G enrichment is also associated with recombination on the X in females; however, the observed C>G enrichment in the strongly male-biased hotspot PAR1 (**Figure 1.5**) supports our conjecture of a male-specific impact of meiotic DSBs on the X, at least in this region.

One possibility is that the enrichment of C>G mutations stems from a switch in the repair machinery late in meiosis [106, 107, 108]; DSBs still not repaired by this stage may be repaired by a more mitotic-like repair pathway, which could potentially be more mutagenic [107, 109]. Notably, the source of the C>G signature found by Jonsson et al., 2017 in specific autosomal regions could also be late repair; indeed, if these areas reflect damage, as the authors surmise, they may only undergo repair later in meiosis. Thus, a shared biochemical pathway may underlie the mutagenic impact of both spontaneous and a subset of meiotic DSBs. Moreover, our results illustrate a subtle sex-specific mutagenic effect of meiotic recombination, whereby the repair of meiotic DSBs on the X specifically in males gives rise to C>G mutations; in that sense, components of the recombination machinery that are involved in late repair of double-strand breaks are sex-specific modifiers of mutation.

1.5 Implications

By comparing the mutation spectrum across different compartments of the genome, we identify putative signatures of sex differences in the germline and plausible biochemical sources of mutagenesis. Notably, we show that replication timing affects the mutation spectrum along the genome and find a mutagenic effect of meiotic recombination that is both sex-specific and X-specific, revealing an appreciable effect of double-strand breaks, both accidental and deliberate, on the mutation spectrum.

Interestingly, our analysis suggests that signatures of sex differences in the germline are likely abundant, but their contributions to the mutation spectrum are subtle relative to those of biochemical processes shared in the two sexes. This finding is hard to reconcile with the idea that male mutations are mostly replication-driven whereas female mutations reflect a large contribution of spontaneous damage, as then we might expect substantially different types of mutations inherited from mothers and fathers. Instead, consistent with a greater role of spontaneous damage and its repair in both male and female germlines [49], our results are most readily explained if male and female mutational mechanisms are overall highly similar, underpinned by the shared mechanisms associated with replication, transcription, methylation, and recombination, and other sources of damage. Subtle differences in the mutation spectrum between males and females could then be expected to arise due to sex-specific rates of damage and repair at different stages in germline development.

These sex differences in germline mutation are modulated by life history traits of males and females. As one example, the proportion of C>G mutations transmitted in

a single generation increases with the age of the mother ([22, 49]. Indeed, even when there are no sex-differences in the biochemical process itself, much of the biochemical machinery that influences mutation must in theory have subtle sex-specific effects, simply because sex-specific life history traits modulate exposure to biochemical influences differently in males and females. Changes in life history traits, or in the frequency of variants associated with sex-specific life history traits over evolutionary time could then change the proportion of particular mutation types and thus alter the mutation spectrum over time. Together with other sex-specific modifiers of mutation, life history traits likely play a role in the evolution of the mutation spectrum not only on autosomes, but also on the X chromosome relative to autosomes.

In this respect, we note that a number of recent studies have shown that the mutation spectrum changes slightly across populations [110, 52, 111, 112]. These findings have largely been attributed to biochemical modifiers of mutation that alter the relative rates of different mutation types by influencing the biochemical process of error/repair over time. Our results highlight that life history traits and other sex-specific modifiers could potentially result in the same kinds of changes in the mutation spectrum and the mutation rate over time. Moreover, parental ages of reproduction explain a large proportion of the observed mutation rate variation among ~1500 individuals at present ([22]). Variants that contribute to sex-specific life history [113, 114] may therefore be a useful starting point to identify genetic sources of inter-individual variation in the mutation rate in humans.

Beyond these insights into mutagenesis, our analysis makes clear that X-autosome comparisons of mutation patterns cannot be taken as directly reflective of germline

sex differences. Though historically comparisons of the sex chromosomes to autosomes have been taken to reflect only the effects of sex, mutation patterns on the X chromosome in fact reflect a convolution of X chromosome specific effects and sex. Taking this point into consideration may help to explain, for instance, why estimates of the male bias in mutation for CpG sites from phylogenetic studies that used X-autosome comparisons are much lower [115] than those obtained directly from male-female differences in de novo mutation data (12, [21, 22]).

1.6 Acknowledgements

We thank Guy Amster, Ziyue Gao, Priya Moorjani, Itsik Pe'er, Jonathan Pritchard, Guy Sella, Arbel Harpak, Felix Wu, and additional members of the Przeworski lab for helpful discussions and/or comments on a draft version of the manuscript and Priya Moorjani, Konrad Karczewski, and Kelley Harris for assistance with gnomAD and SGDP data sets. We are also grateful to comments from three anonymous reviewers on an earlier draft. This work was supported by R01 GM122975 to M.P.

1.7 Supplementary Methods

1.7.1 Delineating the set of variants in the gnomAD dataset

We used publicly available whole genome SNP data from 15,496 individuals compiled and made available by the Genome Aggregation Database (gnomAD), which includes 9,256 Europeans and 4,368 African or African-American individuals [70, 14]. We

restricted our dataset to a set of good quality SNPs that passed the baseline quality filter provided by gnomAD, such that there was at least one individual at each site with a high-quality non-reference genotype: quality-adjusted allele count (AC) > 0, or equivalently, Filter = "Pass", DP > 10, GQ > 20, and AB > 0.2 for heterozygotes. We excluded sites that overlap with indels and CNVs. We retained multi-allelic sites (6.5% and 5% of the data on the autosomes and X respectively). Since our goal was to compare genomic compartments, including those on the X chromosome, we matched the number of X chromosomes and autosomes in our sample by limiting our analysis to the 6,930 female individuals in the sample (using the quality-adjusted female allele counts provided). Additionally, we imposed separate filters on the X chromosome and autosomes so that only variants with an allelic depth within one standard deviation of the mean allelic depth on the X chromosome (13760 +/- 512) and autosomes (13753 +/- 562), respectively, were retained in the sample (only about 2.8% of data from both the X-chromosome and autosomes is lost at this step).

1.7.2 Calculating diversity levels for 96 mutation types

Most variants in the gnomAD dataset are extremely rare: about 64% are singletons and doubletons (i.e., variants seen once or twice in the sample). Only 10% of variants are at frequency 1% or greater (Fig. 1b); their inclusion does not affect our qualitative results, since they are a small subset of the data, and their mutation patterns are largely the same as variants at lower frequencies (**Figure 1.6**); we therefore retain them. The variants in the gnomAD dataset are called with respect to the human

reference (hg19). We instead polarized to the major allele in the full sample of 15,496 individuals, so that the minor allele was treated as derived. At multi-allelic sites, we counted the multiple derived alleles separately as if they had occurred at separate bi-allelic sites with the same major allele. We obtained similar results (**Figure 1.7**) using the ancestral allele and context from the 1000G reconstruction of the ancestral human genome sequence [71]. As is standard (e.g., Alexandrov et al. 2013; Harris 2015; Harris and Pritchard 2017), we extracted the single base pair flanking sequence on each side of the variant position using the portion of the hg19 reference callable in gnomAD to obtain mutations in their trinucleotide context (we substituted the reference allele with the major allele at variant positions to obtain the correct trinucleotide context at these positions; note that the major allele in this sample only differs from the reference allele at 1% of variant positions). We combined mutations in reverse complement classes (for example, the ACG>ATG and CGT>CAT classes were collapsed into the former) to obtain 96 mutation types. To obtain diversity levels for each of the 96 mutation types, we divided the number of segregating sites of a particular type by the number of mutational opportunities at that type of site, where mutational opportunities are defined as sites at which a single change could have given rise to the mutation type under consideration (note that there are three mutational opportunities at each base pair in the genome). By dividing the number of mutations by the number of possible mutations in each genomic compartment, we account for base composition differences between compartments.

1.7.3 Comparing diversity levels between genomic compartments

In comparing a pair of genomic compartments, we took differences in their population genetic properties into account. By normalizing the diversity levels for each mutation type by overall diversity levels for the two compartments, we controlled for the effect of population genetic processes that affect diversity across compartments but are expected to do so evenly across all mutation types, allowing us to isolate differences in the mutation spectrum. This normalization is particularly relevant for comparisons between the X chromosome and autosomes as, for the same sample size, there are more neutral segregating sites expected on autosomes: because the autosomes spend more time in the male germline relative to the X chromosome, they have a higher overall mutation rate, as well as a slightly larger effective population size due to differences in the genealogical process between the X-chromosome and autosomes. Suppose that in an arbitrary genomic compartment "a" with n total mutations ("segregating sites") of all types and n_i mutations of type i, the proportion of mutation type i is n_i/n. If there are d_i potential sites ("mutational opportunities") at which this mutation type could occur, out of d total sites in the compartment, the normalized proportion (or normalized diversity) of mutation type i is:

$$r_{i(a)} = \frac{n_{i(a)}/d_{i(a)}}{n_{(a)}/d_{(a)}} = \frac{\pi_{i(a)}}{\pi_{(a)}}; \qquad n_{(a)} = \sum_i n_{i(a)}, \;\; d_{(a)} = \sum_i d_{i(a)}$$

The relative enrichment of this mutation type in compartment a, compared to another compartment b, is then:

$$R_{i(ab)} = \frac{r_{i(a)}}{r_{i(b)}} = \frac{\pi_{i(a)}/\pi_{(a)}}{\pi_{i(b)}/\pi_{(b)}} = \frac{\frac{n_{i(a)}/d_{i(a)}}{n_{i(b)}/d_{i(b)}}}{\frac{n_{(a)}/d_{(a)}}{n_{(b)}/d_{(b)}}}$$

In particular, when the two compartments under consideration are the X chromosome and autosomes, normalizing by overall diversity allows us to take into account the different population level effects of demography and life history on these compartments, captured in the effective population size (Ne) parameter (under some simplifying assumptions, e.g., no multiple hits), and isolate differences in the mutation spectrum:

$$\frac{\bar{\pi}_{i(X)}/\bar{\pi}_{(X)}}{\bar{\pi}_{i(A)}/\bar{\pi}_{(A)}} = \frac{\mu_{i(X)}.N_{e(X)}/\mu_{(X)}.N_{e(X)}}{\mu_{i(A)}.N_{e(A)}/\mu_{(A)}.N_{e(A)}} = \frac{\mu_{i(X)}/\mu_{(X)}}{\mu_{i(A)}/\mu_{(A)}}$$

where $\bar{\pi}$ is mean diversity and μ denotes the mutation rate. In large samples with recurrent mutations (i.e., repeat mutations or "multiple hits" at the same site), normalizing by overall diversity does not account fully for population level effects, particularly for sites with high mutation rates. In particular, at the highly mutable CpG sites, recurrent mutations are frequently expected in a sample of this size [24, 14]. Because autosomes have a slightly larger effective population size and higher mutation rate compared to the X, we expect more recurrent mutations at these sites on autosomes. Although we include multi-allelic sites in our analysis and can therefore count mutations to three different alleles at a site, since we only observe allele frequencies and not haplotypes, we do not see recurrent mutations of the same type as separate segregating sites. We would consequently under-count recurrent muta-

tions on autosomes, and may observe an apparent enrichment of these types on the X-chromosome. For this reason, differences in the relative diversity at CpG sites on the X chromosome and autosomes must be cautiously interpreted. This concern applies not just to CpG transitions (C>T mutations at CpG sites), which have the highest mutation rate, but potentially also to C>A and C>G mutations at CpG sites, which also have a higher mutation rate than average [20, 116], and for which we observed a substantial decrease in X-enrichment when we counted multiple alleles at a site (**Figure 1.8**). To be conservative, we excluded CpG sites in comparisons between the X chromosome and autosomes. Including them does not change any of our qualitative results, however. The explanation is likely that the difference in effective population size for the X chromosome and autosomes is small, and that the effect of recurrent mutations on the X-autosome comparison is even smaller. This minor effect is mitigated further by including visible multi-allelic sites.

1.7.4 Testing for significant differences in the mutation spectrum between genomic compartments

We tested if mutation type i is distributed the same way in two compartments (a and b) given what would be expected based on the overall distribution of mutations in the two compartments. Effectively, we considered the relationship between the following four ratios for each of 96 mutation types (excluding the 16 CpG types where needed):

$\pi_{i(a)}$	$\pi_{i(b)}$
$\pi_{(a)}$	$\pi_{(b)}$

Or,

$n_{i(a)}/d_{i(a)}$	$n_{i(b)}/d_{i(b)}$
$n_{(a)}/d_{(a)}$	$n_{(b)}/d_{(b)}$

We designated the larger compartment (i.e., with a greater number of mutational opportunities) as the reference compartment (for example, in X-autosome comparisons, the autosomes were used as reference). We assumed that the number of mutations of a particular type in compartment a (the "test" compartment) is binomially distributed with a mean value proportional to the observed diversity for that type in compartment b (the reference compartment), adjusted for overall differences in diversity between the two compartments:

$$n_{i(a)} \sim binom(d_{i(a)}, f_i)$$

$$E(n_{i(a)}) = d_{i(a)} * f_i$$

$$where\ f_i = \pi_{i(b)} * \frac{\pi_{(a)}}{\pi_{(b)}} = n_{i(b)}/d_{i(b)} * \frac{n_{(a)}/d_{(a)}}{n_{(b)}/d_{(b)}}$$

The factor f_i is the expected diversity level for a given type in the test compartment. For each type, we tested if the observed number of mutations in the test compartment differs from the number expected by chance, using the "binom.test" function in R to obtain p-values. Mutational types were considered significantly different in their frequencies in the two compartments if the two-tailed p-value from the binomial

test was below the Bonferroni-corrected 5% significance threshold (=0.05/96). The relative enrichment for each mutation type is given by:

$$\frac{n_{i(a)}}{E(n_{i(a)})} = R_{i(ab)}$$

We also obtained 95% confidence intervals for the relative enrichment of each type from the binomial test. We implicitly assumed that mutations of one type do not impact mutational opportunities of other types. The reason is that because the total number of mutations is much smaller than the number of mutational opportunities, an increase in the number of mutations of one type does not appreciably decrease the mutational opportunities available for other types. We note, however, that the tests for different mutation types are still not fully independent, because the expected diversity for each mutation type in the test compartment depends on the overall relative diversity in the two compartments. If they constitute a large proportion of the total number of mutations, mutation types that are highly significantly enriched in one compartment could influence the null distribution for other mutation types and thus lead to the depletion of these other types in that compartment. In general, we focused on describing the top signals we observed, which are unlikely to be strongly affected by this phenomenon. Nevertheless, to assess the impact of this issue, we implemented a procedure similar to the "forward variable selection procedure" used by Harris and Pritchard (2017). We ranked mutation types by their p-values in an initial set of 96 tests. We then sequentially removed the most significant mutation type and reassessed the other types for significance at each step; mutation types that

reached significance through interactions with other types should drop out. We note that re-generating the ratio of expected diversity in two genomic compartments at each step based on the mutation types remaining in the sample can result in even more significant differences between compartments. Because mutations at CpG sites have the largest sample sizes by far, the largest impact of forward variable selection is observed when CpG transitions are highly enriched in a particular compartment (i.e., in non-CpG islands relative to CpG islands); for this analysis we only highlight the top signals (**Figure 1.16**). Our other analyses remain qualitatively unchanged by forward variable selection, and also by excluding CpG sites. The effect of these procedures on our X-autosome comparison is shown in Supplementary Figure 4. We note that in testing for significant differences in the mutation spectrum between genomic compartments using a binomial test, as described above, we implicitly ignore sampling error in the estimate of diversity of the designated reference compartment; we verified that our results are insensitive to this assumption by using alternative approaches to calculate significance that do not make this assumption, but have other limitations. These are described below: First, we bootstrapped the expected distribution of a particular mutation type in the two compartments using hypergeometric sampling. For a given mutation type i with sample size $n_i (= n_{i(a)} + n_{i(b)})$, we generated a random variable k_i for the number of draws in compartment a when n_i mutations were drawn without replacement from the pool of all mutations in both compartments, $n = n_{(a)} + n_{(b)}$ with $n_{(a)}$ "successes" in compartment a, i.e., $k_i \sim Hyper(n, n_{(a)}, n_i)$. The expected distribution of the ratio of mutations of type i in the two compartments, $\frac{k_i}{n_i - k_i}$, was obtained using 10,000 such trials. We calculated a p-value using

the rank of the observed relative diversity (adjusted by the ratio of total mutational opportunities of all types in the two compartments) in the expected distribution. In this approach, we allowed for uncertainty in the number of mutations of a particular type in both compartments, but held constant the total diversity in each compartment. Second, we adjusted the number of mutations in one of the compartments by the ratio of overall diversity in the two compartments, and then applied a chi-squared test to the 2x2 contingency table (shown below) of the mutations ("successes") and the remaining mutational opportunities ("failures") by compartment. Third, we fitted the same 2x2 contingency table using a binomial glm (with the compartment as a covariate).

$n_{i(a)}/p$	$n_{i(b)}$
$d_{i(a)} - (n_{i(a)}/p)$	$d_{i(b)} - n_{i(b)}$

where $p = \dfrac{n_{(a)}/d_{(a)}}{n_{(b)}/d_{(b)}}$

We compared these methods for three analyses: X vs Autosome, PAR vs Autosome, and X-A matched for replication timing (Supplementary Methods Tables 1-3). In all cases, the same significant mutation types (and the same effect sizes) stand out; in other words, the results are the same, regardless of the approach. The reason is likely that the reference compartment is always sufficiently large for sampling error to be small.

1.7.5 Comparing the X chromosome and autosomes: additional considerations

In comparing compartments on the X chromosome and autosomes, we excluded the pseudo-autosomal region unless otherwise specified, since sex-specific properties differ between the PAR and the rest of the X chromosome. We considered additional population-level properties that may affect the mutation spectrum differently for the X and autosomes. In accordance with our prior expectation that biased gene conversion should have a negligible effect on variants at very low frequencies, we note no clear patterns of X-autosome differences for mutation types that are subject to biased gene conversion. Similarly, because so many variants in the sample are rare and thus young, we expect very little effect of either direct or linked selection on the mutation spectrum a priori. Thus, we interpret these mutation patterns as reflecting real and largely neutral differences between X and autosomes. We also checked for differences in genotyping error rates between the X chromosome and autosomes. Because the variant quality (QUAL) variable in the dataset is jointly calculated based on males and females in the sample, the reported quality of variants on the X chromosome is expected to be slightly lower. Nevertheless, the distribution of variant quality is almost identical for the X chromosome and autosome (**Figure 1.11**), and any small differences are likely further lessened because we used only the female subset of the data. To rule out a potential interaction of error rate by mutation type and compartment, we compared the genotype qualities and read depths for C>G and C>A mutation types in compartments across which we found the distribution of these

types to differ. The average genotype quality and read depth is high for all mutation types, and while there may potentially be small differences in error rates between different types of mutation, these do not seem to differ across compartments in ways that would affect our inference. For instance, average genotype quality and read depth for C>G mutations is similarly distributed to all other mutation types combined both in the PAR and Autosomes, while a strong enrichment of these types is seen only in the PAR. We similarly did not observe a notable interaction of C>A quality metrics with early and late replicating compartments; these cases are illustrated in **Figure 1.12**. We further replicated our analysis using two datasets (Uk10k and SGDP) (The UK10K Consortium, 2015; Mallick et al., 2016) sequenced independently, with varying levels of coverage (**Figure 1.10**). In order to validate signals observed in the gnomAD data, we conducted a similar analysis using publicly available data from the Simons Genome Diversity Project (SGDP) [117]. We considered only individuals with ID beginning with "S" (as this subset was sequenced PCR-free and processed using a consistent approach) and used filter level 1 (recommended as optimal for SNP discovery in Mallick et al. 2016) to obtain a variant set of high quality. The SGDP variants were polarized with respect to the major allele in the full SGDP sample of "S" individuals. We limited our analysis to female individuals (100/256 individuals in the dataset). Individual-level alternate allele counts at each variant position, reported as 0, 1 (heterozygous), or 2 (homozygous), were summed over the 100 individuals to obtain allele counts in the sample; no multi-allelic sites were seen in this sample. Sites with >50% missing data were excluded. The major allele in the SGDP dataset is 99% correlated with the gnomAD major allele when only variants at matched

positions are considered. To estimate the mutational opportunities for each type of site in this dataset, we combined accessible regions from 11 individuals (five with predominantly European and six with African ancestry) and used the hg19 reference to obtain trinucleotide context. We also replicated our analysis using variants in the ALSPAC and TwinsUK subsets of the UK10K dataset (The UK10K Consortium, 2015). We limited our analysis to the 2,793 females. We polarized to the major allele in the UK10K sample, and applied quality thresholds similar to those in the gnomAD analysis. We excluded multi-allelic sites from this sample. Since accessible regions were unavailable for this dataset, we used the accessible regions obtained for the SGDP dataset, obtained as described above.

1.7.6 Obtaining data for the distribution of genomic features

To investigate the association of biochemical features with mutational patterns in the germline, we would ideally consider the distribution of these features in germline tissue. In cases where we were limited to data from somatic cell lines, where possible we focused on genomic features that are known to have stable or broadly similar distributions across tissues-types, as these are more likely to be comparable between the soma and germline. We downloaded replication timing data from two sources: data for LCL lines was obtained from Koren et al., 2012, and data for three human embryonic stem cell lines (H1, H7, and H9), produced as part of the ENCODE project, was downloaded from the UCSC browser. Replication timing data are reported as a

standardized score with negative scores representing later replication. To check for systematic differences in the observed distributions of replication timing between the two studies, we also downloaded data for the cell line most similar to LCL available from ENCODE (GM12878); for these cells, the distribution of replication timing was almost identical to the LCL data from Koren et al., 2012, suggesting that differences in the distributions of replication timing between the LCL and embryonic stem cell lines are not due to methodological differences, and likely reflect real biological differences between cell lines. For autosomes, average replication timing is largely consistent across different cell lines, both at the chromosomal level and at the 1Mb scale (**Figure 1.13**) [27, 73]. Since methylation is expected to be highly variable across tissues and may well differ substantially between the germline and soma, we used CpG islands as a binary proxy for methylation: CpG islands are hypomethylated relative to the rest of the genome, across tissue-types [74, 75]. We obtained the X-inactivation status ("inactive", "escape", "variable", "unknown") of genic regions on the X chromosome from Tukiainen et al. 2017; this consensus status for 683 genes on the X chromosome is based on combined information from multiple sources and experimental approaches, across tissue-types. To be consistent with the Tukiainen et al. 2017 study, we used Gencode v19 coordinates and annotations for genic regions on the X chromosome and autosomes. The small number of regions of overlap between genes that were classified as both "escape" and "inactive" were excluded from the analysis. DMC1 ChIP-seq signal intensity on the X chromosome and autosomes, measured in spermatocytes, was obtained from Pratto et al., 2014. DMC1 hotspots were defined as 1 kb regions around the midpoint of hotspots identified by Pratto et al., 2014. We used hotspots

and signal intensity values for the "AA2" individual; using average intensities and the union of hotspots from all four individuals with PRDM9 alleles A and B does not alter our qualitative results. DMC1 hotspots were grouped as weak (DMC1 signal intensity 1-50), intermediate (signal intensity 50-150), and strong (signal intensity >150). Because the X chromosome has a disproportionate number of very strong DMC1 hotspots, we chose these criteria to obtain similar average hotspot intensities on the X chromosome and autosomes in the first three bins, with the outliers in the fourth; varying these thresholds does not alter our qualitative conclusions. We obtained the list of autosomal regions enriched for clustered C>G mutations from Jónsson et al., 2017. Finally, we used the female standardized recombination map [118] to define female hotspots on autosomes (following Kong et al. 2010, windows with recombination rates greater than 10-fold the genome average were considered hotspots; increasing this threshold does not alter our qualitative conclusions).

1.7.7 Testing the effect of replication timing and other genomic features on the autosomal mutation spectrum

For autosomes, average replication timing is largely consistent across cell lines, both at the chromosomal level and at the 1Mb scale (**Figure 1.13**) [27, 73]. We compared the mutation spectrum in autosomal regions that are early or late replicating in LCL (Fig. 2a) and H9-hESC (**Figure 1.14**) cell lines. Regions were defined as early replicating if the replication timing score was greater than or equal to 0.5, and late replicating if it was less than or equal to -0.5 (the results remain qualitatively

unchanged if these thresholds are varied). We aggregated replication timing data per 1 Mb window and per chromosome (using the bedtools map function). While the choice of scale is somewhat arbitrary, averaging replication timing on the 1 Mb scale is relatively lossless (**Figure 1.13**), and this scale has been used in other studies of replication timing [65, 27]. In each 1 Mb window, we obtained the enrichment of each of six broad mutational classes (C>A, C>G, C>T, T>A, T>C, T>G) relative to all mutation types, and relative to all windows taken together. We excluded a small number of windows in which the total number of mutations was outside the range of two standard deviations from the mean. For the chromosome level analysis, for each autosome, we obtained the enrichment of each of the six broad mutational classes relative to all mutation types, and relative to all other autosomes taken together. Averaging replication timing on entire chromosomes is useful because it allows us to place the effect of replication timing on the X chromosome in context. To assess the impact of other genomic features, we compared the autosomal mutation spectrum in regions that lie within and outside CpG islands, and regions that lie in regions of clustered de novo mutations thought to be due to double-strand break damage (**Figure 1.16**). We note that many genomic features are correlated. As one example, hypomethylated CpG islands tend to colocalize with early replicating gene regulatory regions; we consider the effect of this interaction on the mutation spectrum (**Figure 1.15**).

1.7.8 Controlling for genomic features on the X-chromosome

For the X chromosome, the average difference in replication timing between cell lines is thought to reflect X-inactivation status, which differs by cellular genotype and the level of differentiation, and can be heterogenous in a cell population [90, 27, 92, 73, 91]. For instance, the LCL line is a female (XX) cell line with one stably inactivated X chromosome, consistent with the significantly later replication of the X on average (Supplementary Figure 8). The H1 embryonic cell line is male (XY), whereas the H7 line is XX but with two active X chromosomes, and in both these cases the X replicates on average at the same time as autosomes (**Figure 1.13**). We did not observe late replication on the X chromosome for the female H9 cell line (**Figure 1.13**), which is thought to have one inactive X chromosome. The reason may be that this particular line is derived from 5-day old female embryos: since the pre-implantation embryo undergoes global demethylation and is hypomethylated around day 5 post-fertilization [60, 59], we do not necessarily expect to see an effect of X-inactivation in this cell line. Thus, we did not have data for human embryonic stem cells where one X chromosome is stably inactive; however, the differences in X-inactivation for human embryonic cells in various states of differentiation is consistent with X-inactivation status changing over time in the germline, and on average lying somewhere in between that observed in the LCL cell lines and the human embryonic stem cell lines. More generally, these patterns support the notion that X-inactivation and replication timing on the X chromosome are highly correlated. We controlled

for X-inactivation on the transcribed X chromosome using the X-inactivation status. Complementary to this approach, and to ensure that the observed impact of X-inactivation was not an artifact of a potential mis-classification of genes in the active and inactive categories, we controlled for replication timing on the X chromosome. To match the average replication timing on the X and autosomes, we considered regions on the X chromosome and autosomes that have replication timing scores of greater than or equal to -0.5 and less than or equal to 0.5 in the LCL cell line (**Figure 1.18**). The total length of callable regions (in gnomAD) with this property is about 700 Mb on autosomes and 50 Mb on the X. On average, variants in these regions on both the X and autosomes have a similar (approximately zero) mean score for replication timing. In matching the mean, we implicitly assumed that the effects of replication timing on the mutation spectrum were roughly linear (there are some non-linear effects on the X chromosome in regions of extremely late replication timing). Changing the threshold does not alter our qualitative results as long as the mean replication timing on the X chromosome and autosomes is close. We also considered shared mutational patterns in early and late replicating regions of the X relative to autosomes (**Figure 1.19**). Since the mean replication timing on the X is -0.75, we considered late replicating regions to be those with a replication timing score $<$ -1.25, and early replicating regions, $>$ -0.25; changing these thresholds changes the power of our comparison but not the qualitative results. We ignored effects of differences in CpG methylation between the X and autosome since we excluded CpG sites in the X-autosome comparison. We assumed that any additional small effects of methylation at other sites were controlled for indirectly by controlling for replication

timing and/or X-inactivation. For the pseudoautosomal compartment, we did not control for genomic features, since it does not undergo inactivation and replicates early and thus, the mutation spectrum in this compartment should not be affected by these features. All annotation sources are listed in (**Table 1.1**).

1.8 Supplementary Tables and Figures

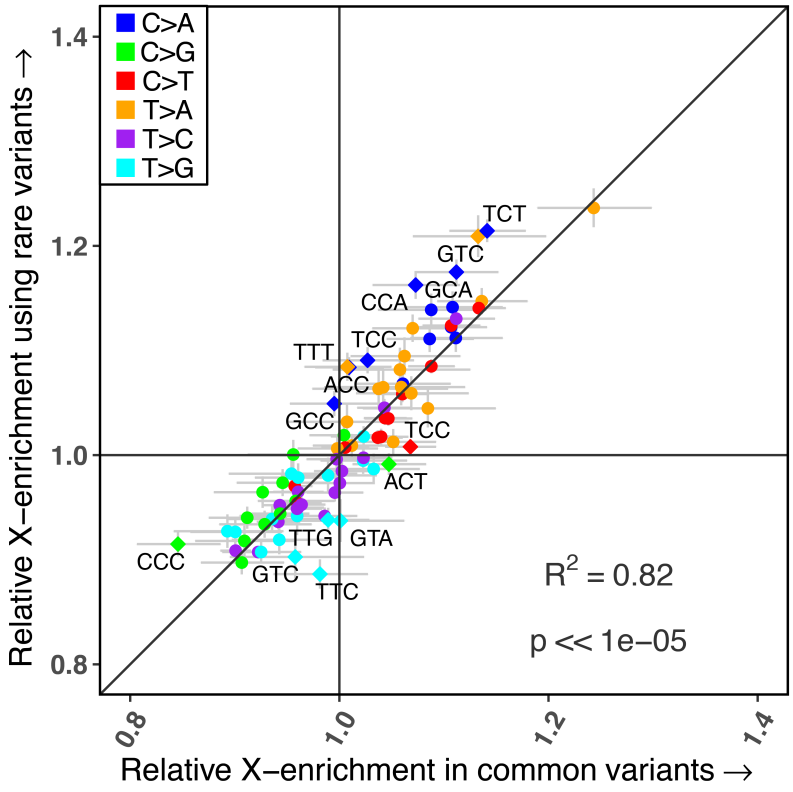

Figure 1.6: The enrichment of mutation types on the X chromosome, relative to autosomes, for rare variants (allele counts 1-5) versus common variants (variants at frequency 1% or greater) in the sample. This analysis excludes CpG sites and the pseudoautosomal region. Diamonds represent mutation types that differ by 5% in their mean fold enrichment between the two comparisons. The black diagonal represents the x=y line. For each estimate, the 95% confidence interval from a binomial test for X-enrichment among common and rare variants, respectively, is represented by the grey horizontal and vertical bars.

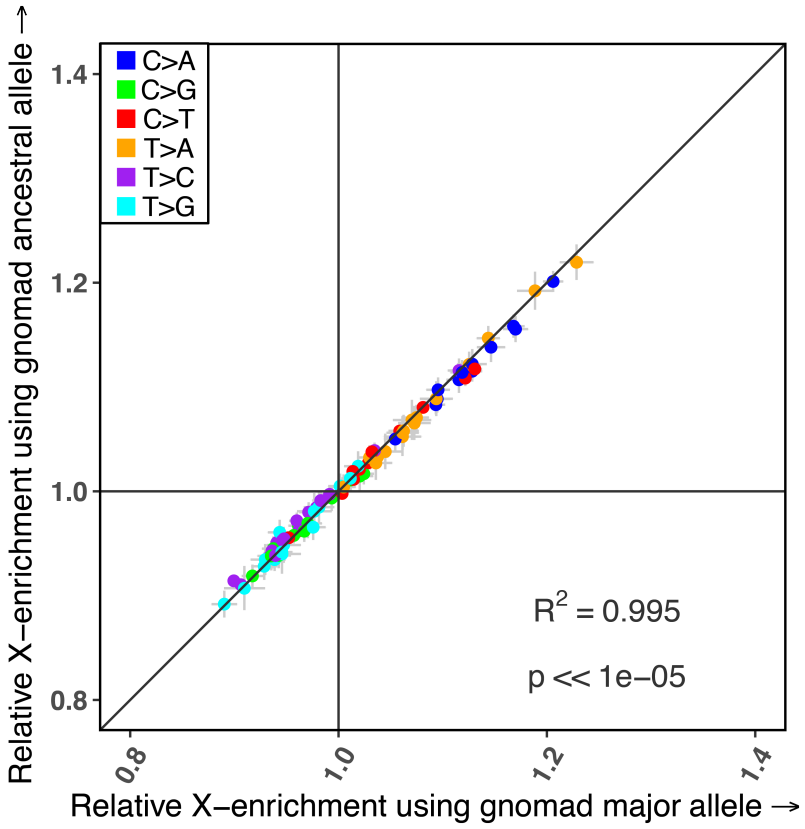

Figure 1.7: The enrichment of mutation types on the X chromosome, relative to autosomes, using the major allele in the sample versus the ancestral allele from the human chimp ancestral sequence. This analysis excludes CpG sites and multi-allelic sites. The black diagonal represents the x=y line. For each estimate, the 95% confidence interval from a binomial test for enrichment on the X chromosome relative to autosomes, using the major allele or the human chimp ancestral allele, respectively, is represented by the grey horizontal and vertical bars.

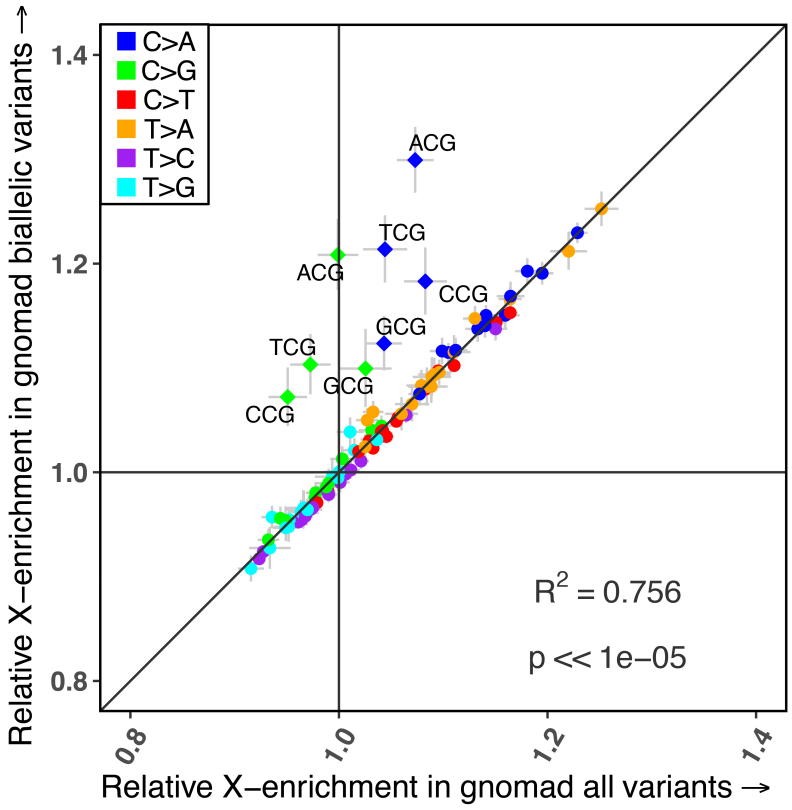

Figure 1.8: The enrichment of mutation types on the X chromosome, relative to autosomes, in all variants including multi-allelic sites, or bi-allelic sites only. Diamonds represent mutation types that differ by 5% in their mean fold enrichment between the two comparisons. The black diagonal represents the x=y line. For each estimate, the 95% confidence interval from a binomial test for enrichment on the X chromosome relative to autosomes, using all sites or only bi-allelic sites, respectively, is represented by the grey horizontal and vertical bars.

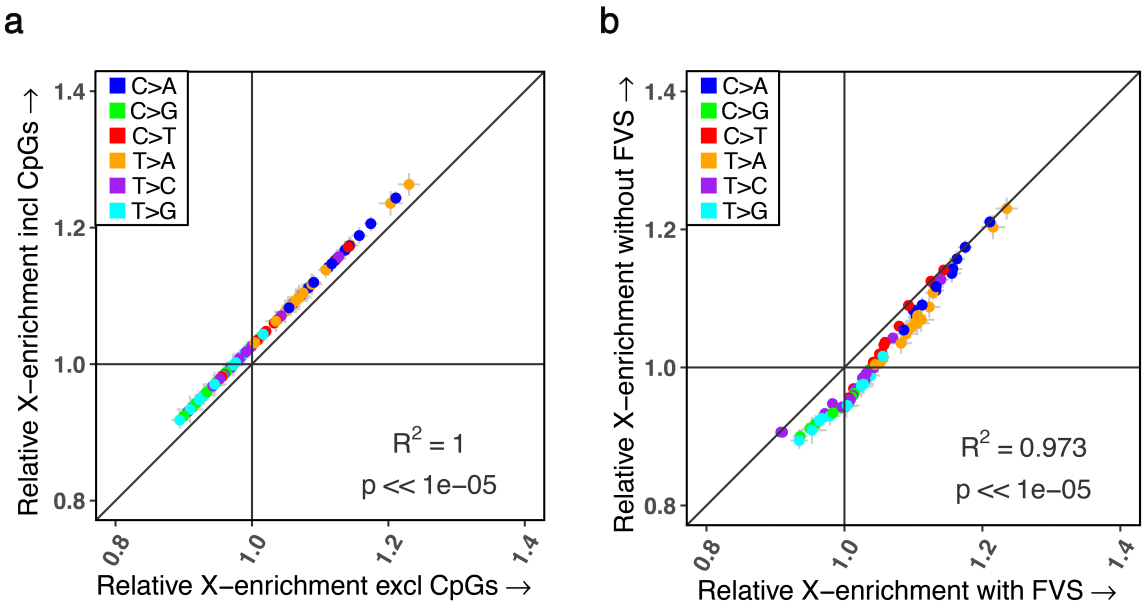

Figure 1.9: The effect of forward variable selection and excluding CpG sites on the X-autosome comparison. The black diagonal represents the x=y line. For each estimate, the 95% binomial confidence intervals are represented by the grey horizontal and vertical bars. (a) The enrichment of mutation types on the X chromosome, relative to autosomes, when all 96 types are considered, versus when CpG sites are excluded. (b) The enrichment of mutation types on the X chromosome, relative to autosomes, with and without the forward variable selection procedure (see Supplementary Methods).

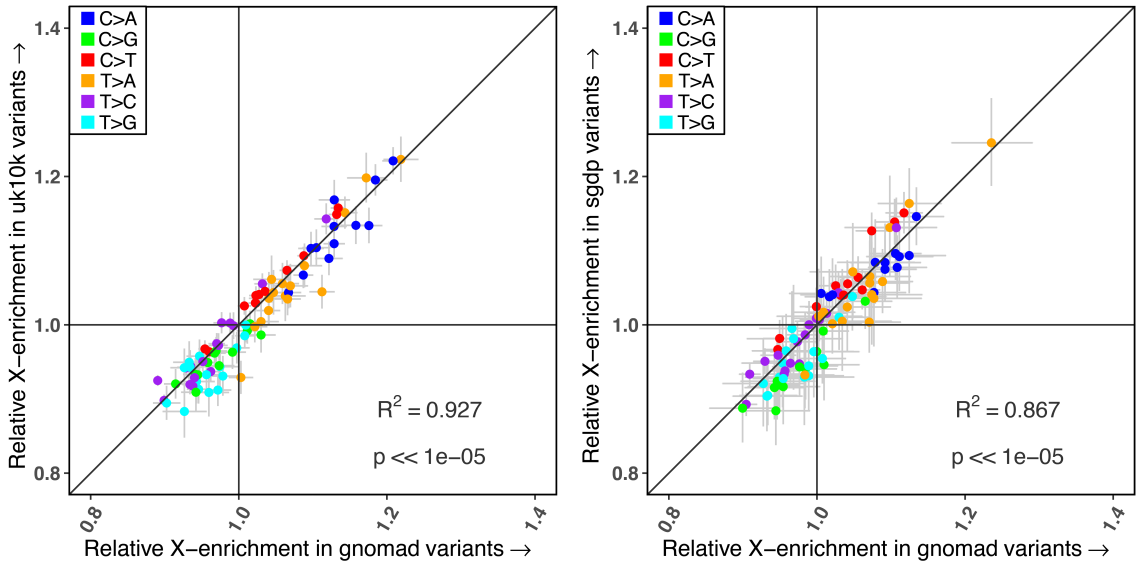

Figure 1.10: Comparison of the X-Autosome mutation spectrum in gnomAD with the UK10K and SGDP datasets. The gnomAD dataset was down-sampled to match the comparison dataset in each case. These analyses exclude CpG sites and multi-allelic sites. The black diagonal represents the x=y line. For each estimate, the 95% confidence interval from a binomial test for X-enrichment relative to autosomes in gnomAD and the UK10K or SGDP datasets, respectively, is represented by the grey horizontal and vertical bars. (a) The enrichment of mutation types on the X chromosome, relative to autosomes, in gnomAD versus UK10K. (b) The enrichment of mutation types on the X chromosome, relative to autosomes, in gnomAD versus SGDP.

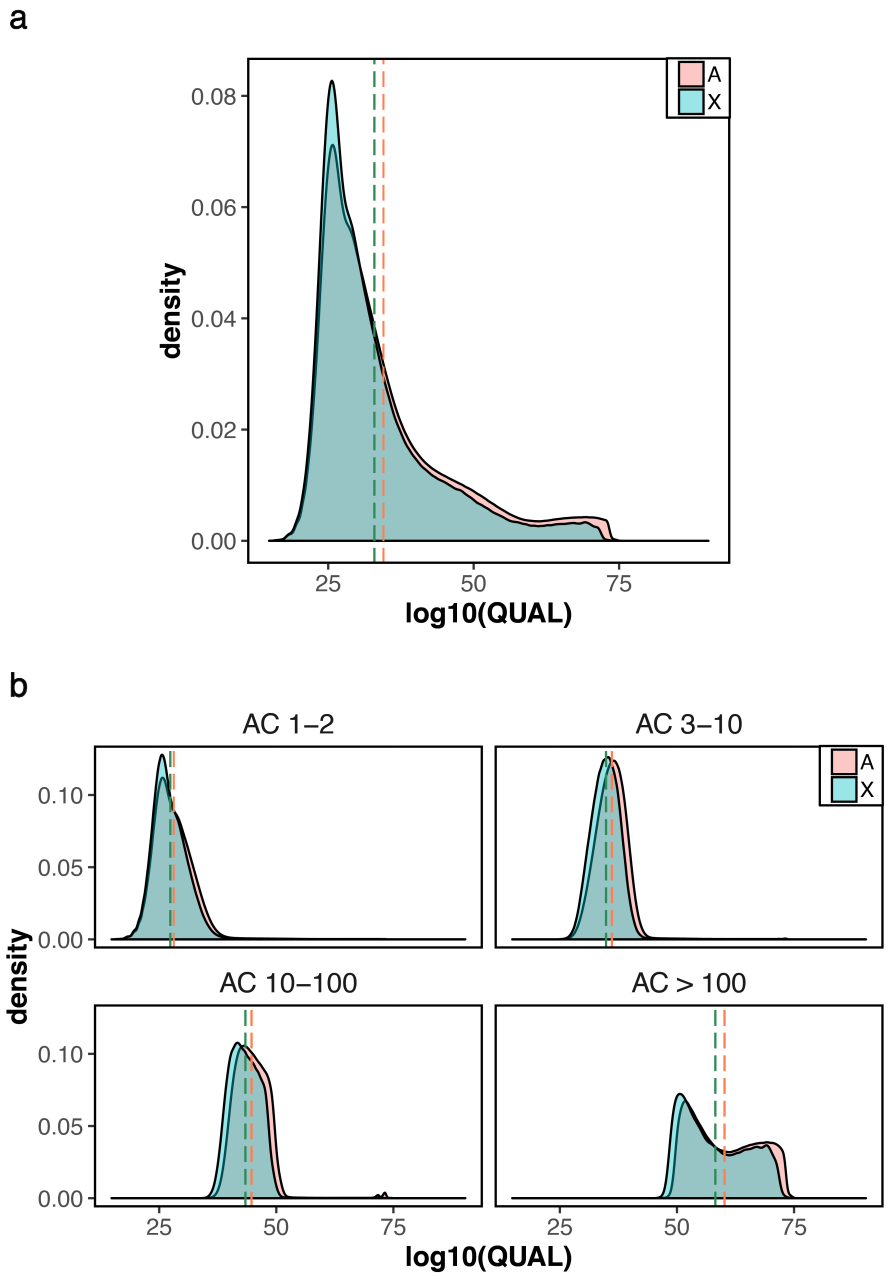

Figure 1.11: (a) Distribution of Variant Quality on the X chromosome and Autosomes (b) Distribution of Variant Quality on the X chromosome and Autosomes by frequency bin. Note that this reported measure of variant quality is based on the full sample with males and females and might be expected to be slightly lower on the X chromosome (see Supplementary Methods)

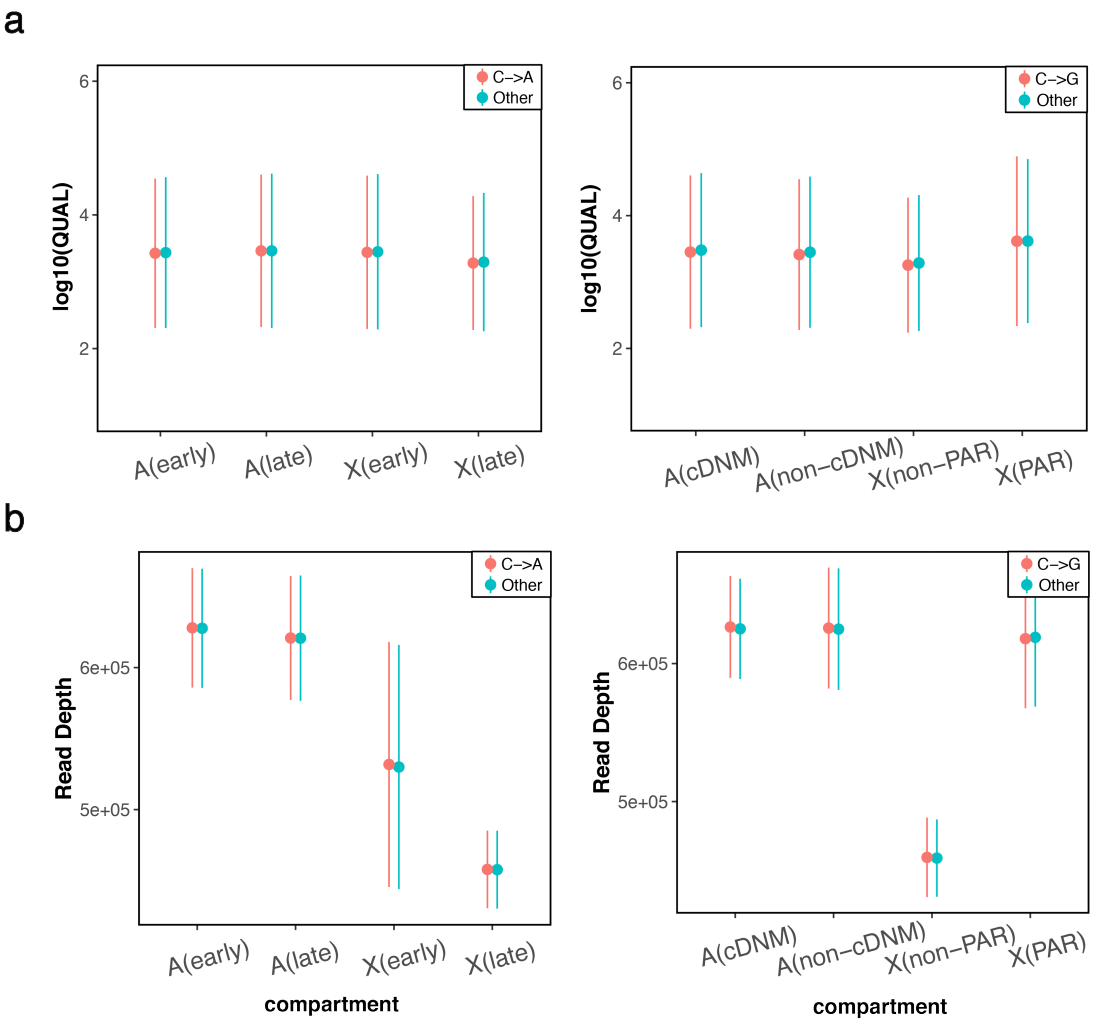

Figure 1.12: Average genotype quality and read depth by mutation type and compartment. Note that these reported measures of variant quality are based on the full sample with males and females and might be expected to be slightly lower on the X chromosome. (a) Genotype quality for C>A mutations types in X-Autosome compartments with differences in replication timing, and for C>G versus all other mutation types in the PAR and other relevant X-Autosome compartments. (b) Read Depth for C>A mutations types in X-Autosome compartments with differences in replication timing, and for C>G versus all other mutation types in the PAR and other relevant X-Autosome compartments.

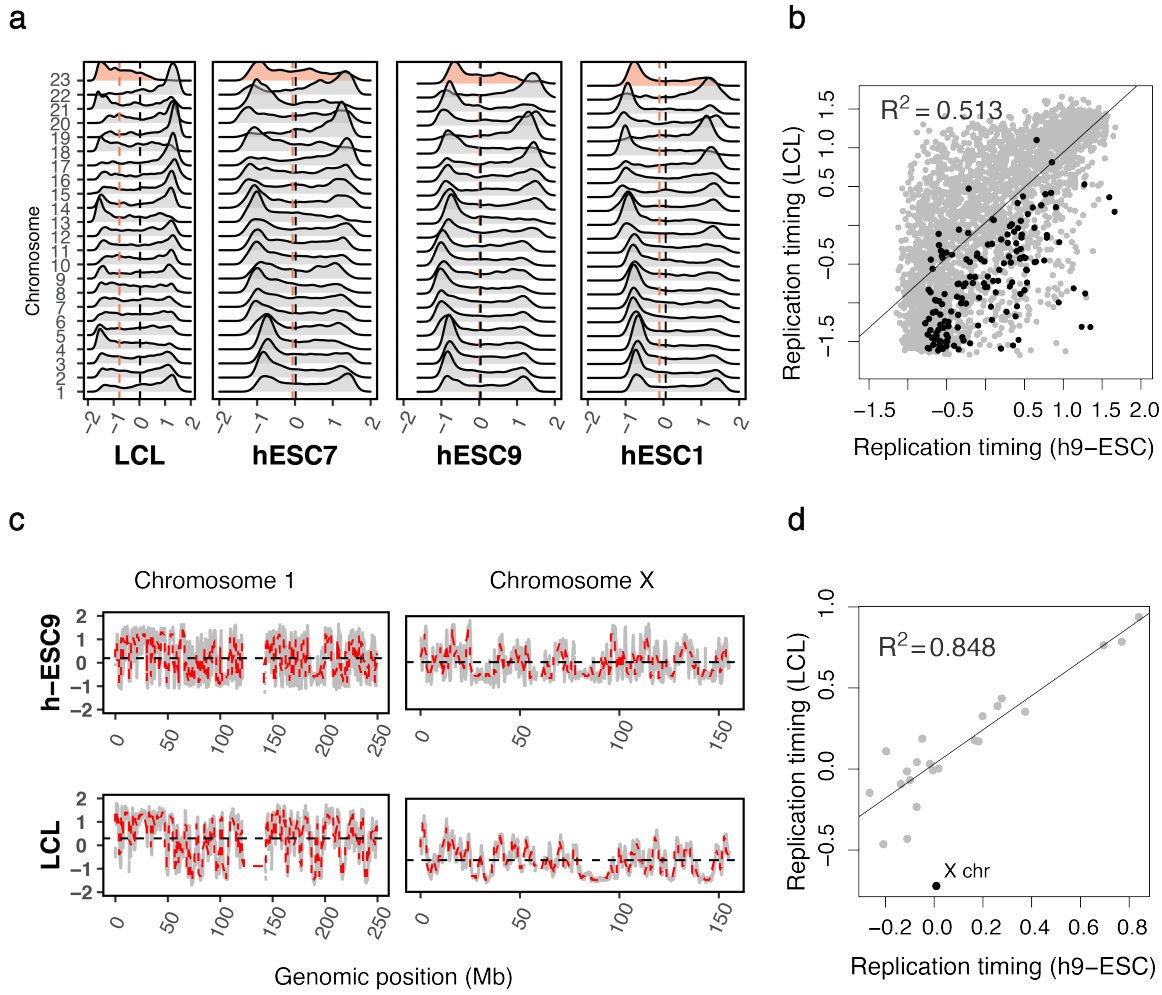

Figure 1.13: Variation in replication timing scores by cell type. Positive values indicate early replication. (a) The distribution of replication timing scores for LCL and human embryonic stem cell lines (hESC1, hESC7, hESC9). The X chromosome is shaded in red. The average scores on autosomes and on the X are denoted by the red and black vertical lines, respectively. (b) Mean replication timing for 1 Mb windows for the LCL and hESC9 cell types. Autosomal windows are shown in grey; windows on the X chromosome have been overlaid in black. (c) Fine scale replication timing for chromosomes 1 and X. Grey points reflect raw replication timing data in bins of approximately 100 bp. The dashed red lines reflect averages over 1 Mb windows. The horizontal black line indicates the chromosome-level mean replication timing. (d) Mean replication timing at the chromosomal level for the LCL and hESC9 cell types.

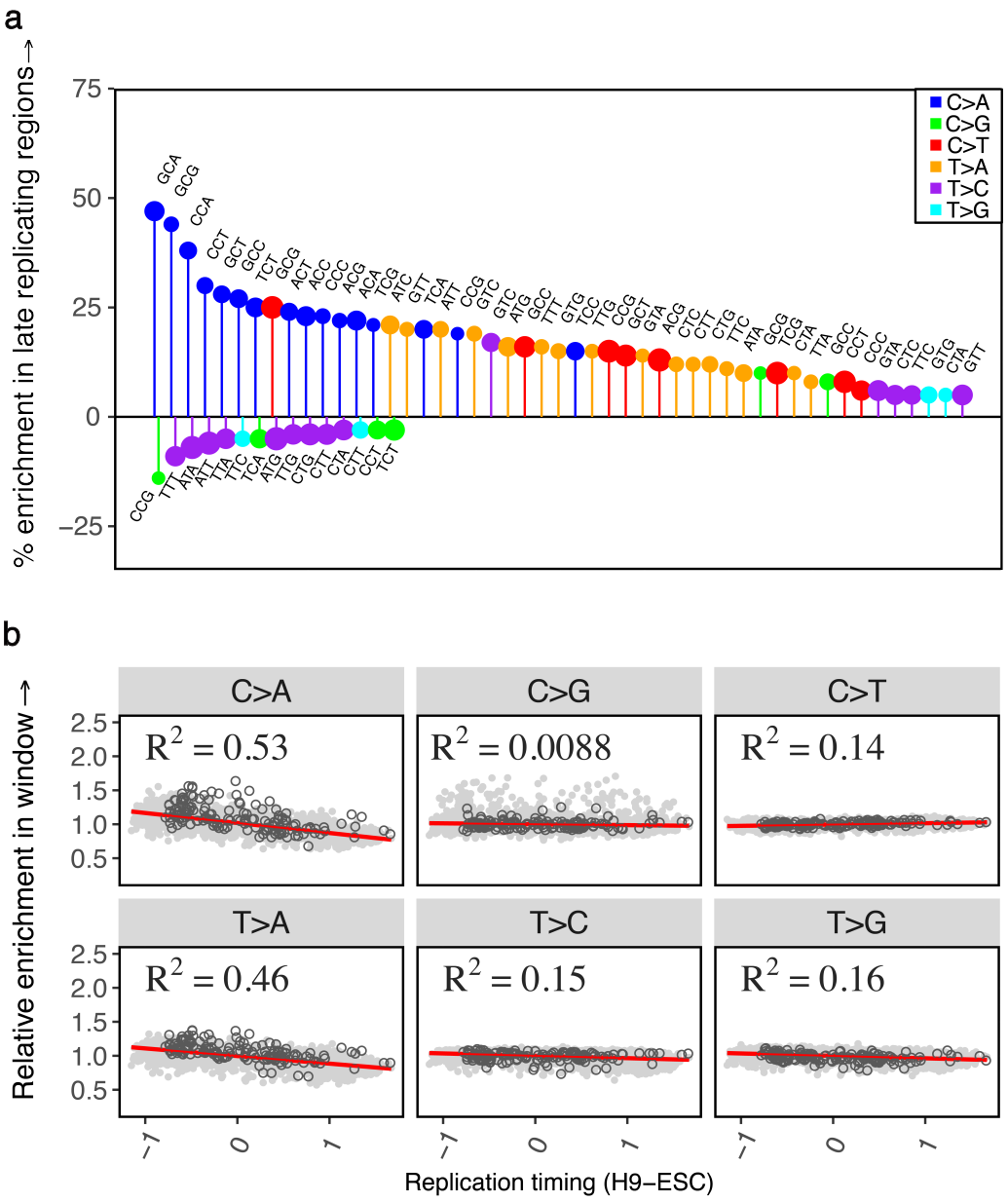

Figure 1.14: The effect of replication timing on the mutation spectrum using the hESC9 cell type. (a) Comparison of the spectrum of 96 mutation types in late replicating autosomal regions relative to early replicating autosomal regions. Only significant differences are shown. Positive and negative effects are ranked separately in order of effect size from left to right. Late replicating regions are defined as having a replication timing score <= -0.5 and early >= 0.5. (b) The relative enrichment of six mutational classes in 1Mb windows relative to all autosomal windows combined, ordered by the mean replication timing in 1Mb windows. Positive x-values indicate early replication. Windows on autosomes are shown in solid grey circles; windows on the X chromosome have been overlaid in black open circles.

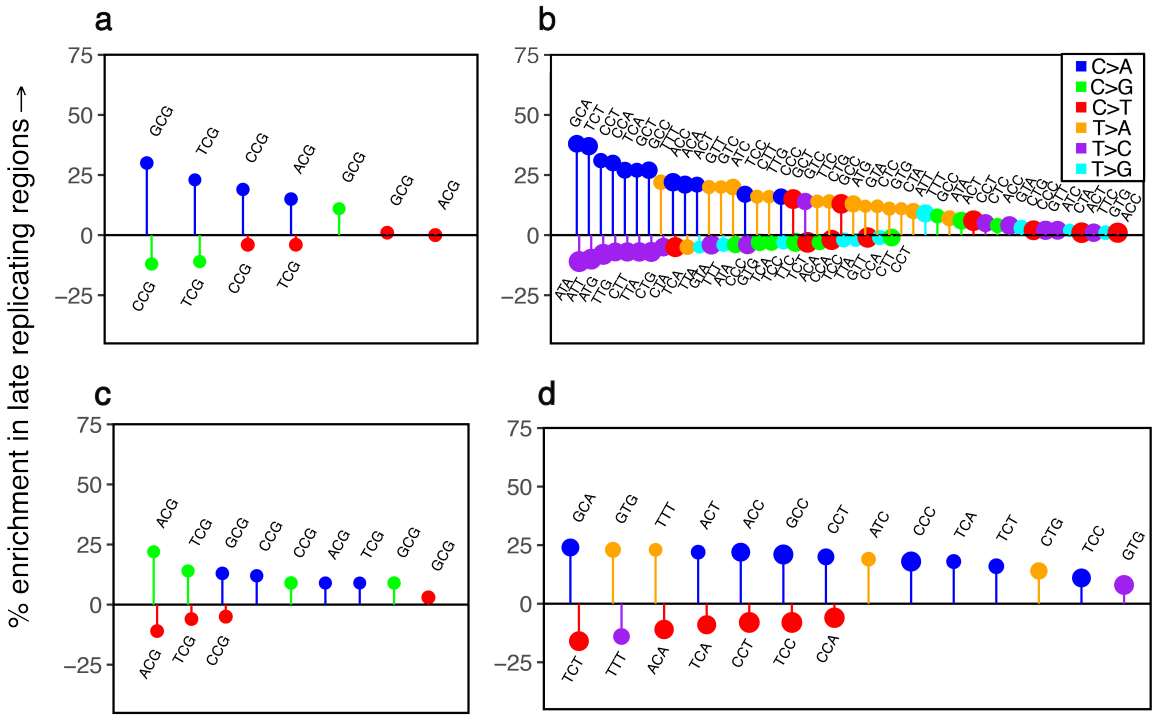

Figure 1.15: The effects of methylation and replication timing on the mutation spectrum. Only significant differences are shown. Positive and negative effects are ranked separately in order of effect size from left to right. CpG islands are used as a proxy for regions of hypomethylation. Late replicating regions are defined as having a replication timing score <= -0.5 and early >= 0.5. (a) Comparison of the mutation spectrum at CpG sites in late versus early replicating regions outside CpG islands. (b) Comparison of the mutation spectrum at non-CpG sites in late versus early replicating regions outside CpG islands. (c) Comparison of the mutation spectrum at CpG sites in late versus early replicating regions inside CpG islands. (d) Comparison of the mutation spectrum at non-CpG sites in late versus early replicating regions inside CpG islands.

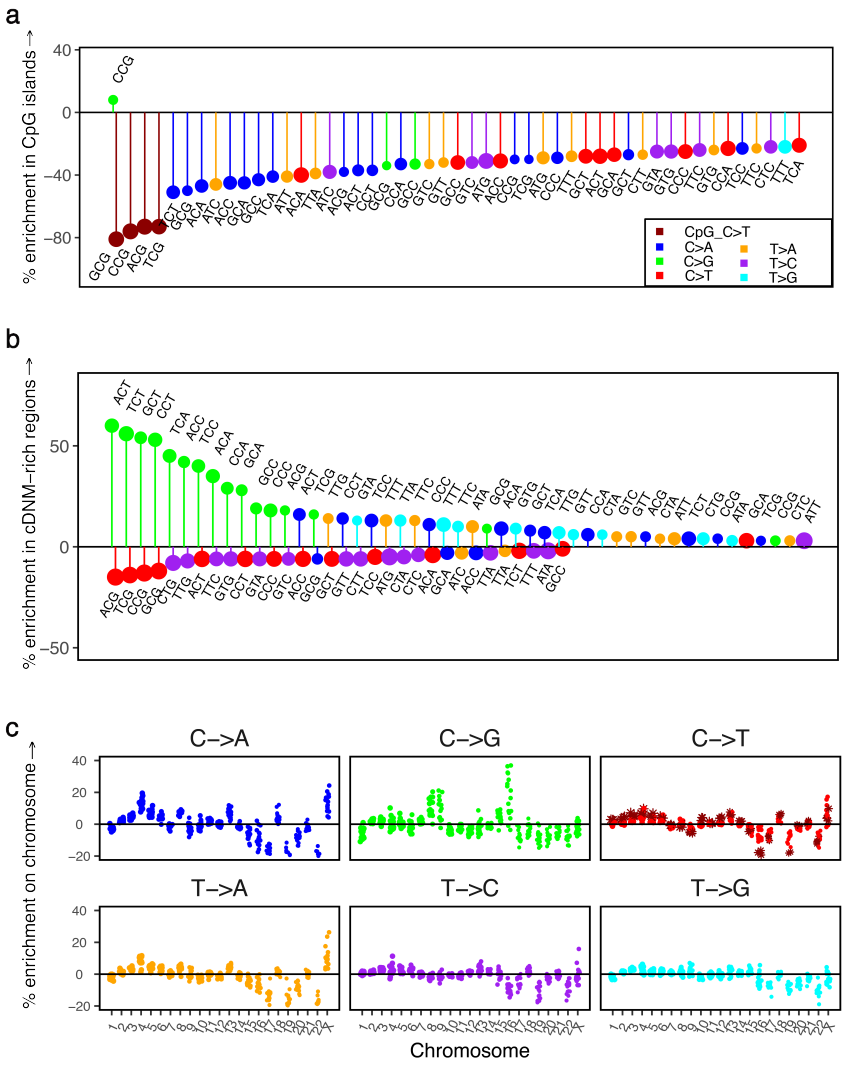

Figure 1.16: The effect of biochemical features on the mutation spectrum. (a) Comparison of the spectrum of 96 mutation types in autosomal regions in CpG islands relative to autosomal regions outside CpG islands. Positive and negative effects are ranked separately in order of effect size from left to right; only the top 50 significant positive and negative effects are shown for legibility. The size of the circle reflects the number of mutations of that type. CpG transitions are labeled in dark red. (b) Comparison of the spectrum of 96 mutation types in autosomal regions identified as rich in clustered de novo mutations (cDNMs) by Jónsson et al., 2017, relative to other autosomal regions. Only significant differences are shown. Positive and negative effects are ranked separately in order of effect size from left to right. The size of the circle reflects the number of mutations of that type. (c) Variation in the mutation spectrum for individual chromosomes. The enrichment level is shown for individual autosomes relative to all other autosomes combined and, on the X, relative to all autosomes combined; each point reflects one of 16 trinucleotide contexts for a particular mutational class. CpG transitions are labeled in dark red.

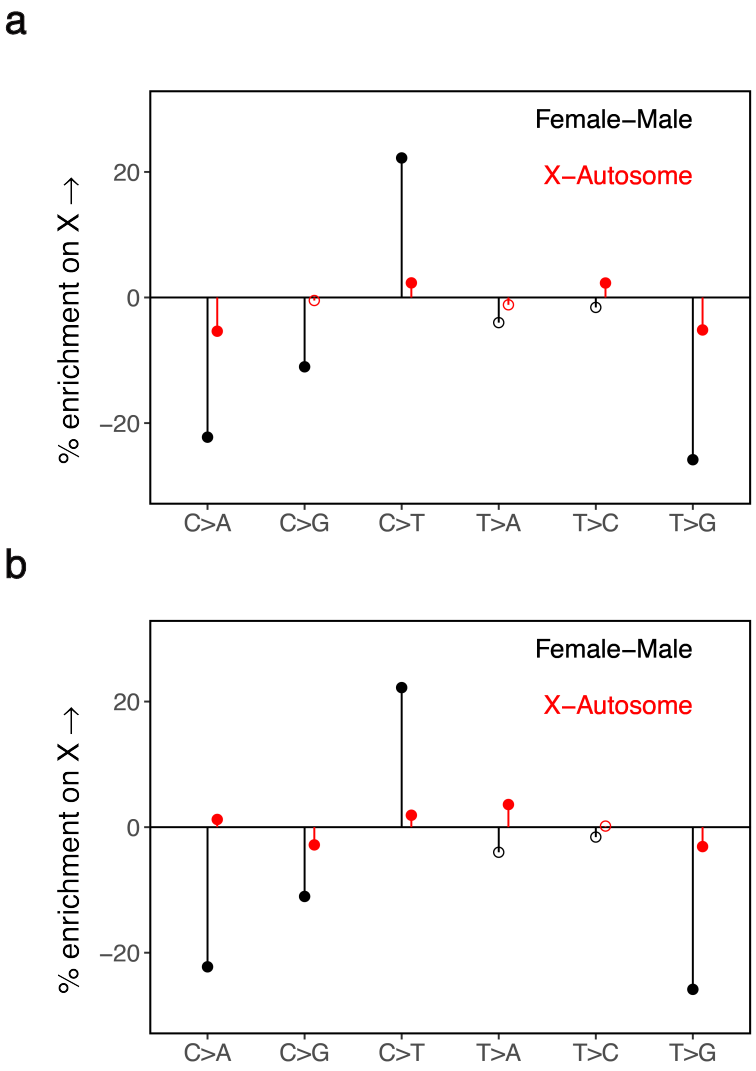

Figure 1.17: The mutation spectrum in the active and inactive genic regions of the X chromosome relative to genic regions in autosomes. These X-Autosome comparisons exclude all CpG sites and the pseudoautosomal region (PAR). (a) The spectrum of six mutational classes in the active genic regions of the X chromosome relative to autosomes (in red), compared to known male female differences from Jónsson et al., 2017 (in black). Solid points are statistically significant differences at the 5% level, accounting for multiple tests. (b) The X-Autosome mutation spectrum in inactive genic regions of the X relative to autosomes (in red) compared to known male female differences from Jónsson et al., 2017 (in black). Solid points are statistically significant differences at the 5% level, accounting for multiple tests.

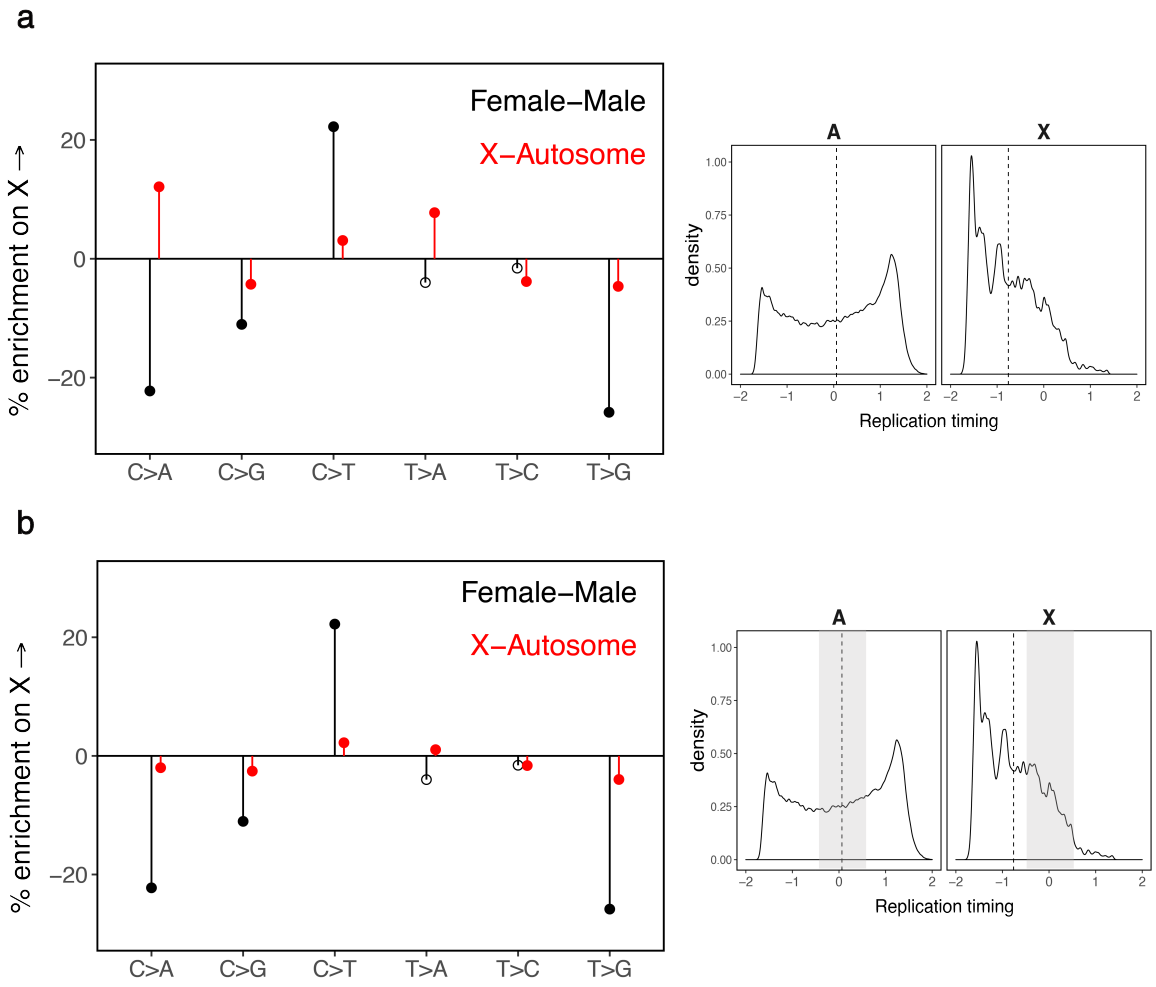

Figure 1.18: The mutation spectrum on the X chromosomes and autosomes with and without matching for average replication timing. These X-Autosome comparisons exclude all CpG sites and the pseudoautosomal region (PAR). The distribution of replication timing for the X and autosomes is shown in the right panels, with the mean replication timing on the X and autosomes represented by dashed vertical black lines. (a) The X-Autosome mutation spectrum (in red), unadjusted for replication timing differences, compared to known male female differences from Jónsson et al., 2017 (in black). Solid points are statistically significant differences at the 5% level, accounting for multiple tests. (b) The X-Autosome mutation spectrum of the X and autosome matched for average replication timing, in red, compared to known male female differences from Jónsson et al., 2017 (in black). Solid points are statistically significant differences at the 5% level, accounting for multiple tests. Note that the left panel is a duplicate of Fig. 4b, to enable a direct comparison between the results with and without matching for replication timing. The shaded regions in the right panel indicate the range of replication timing (between -0.5 and 0.5) used in this analysis.

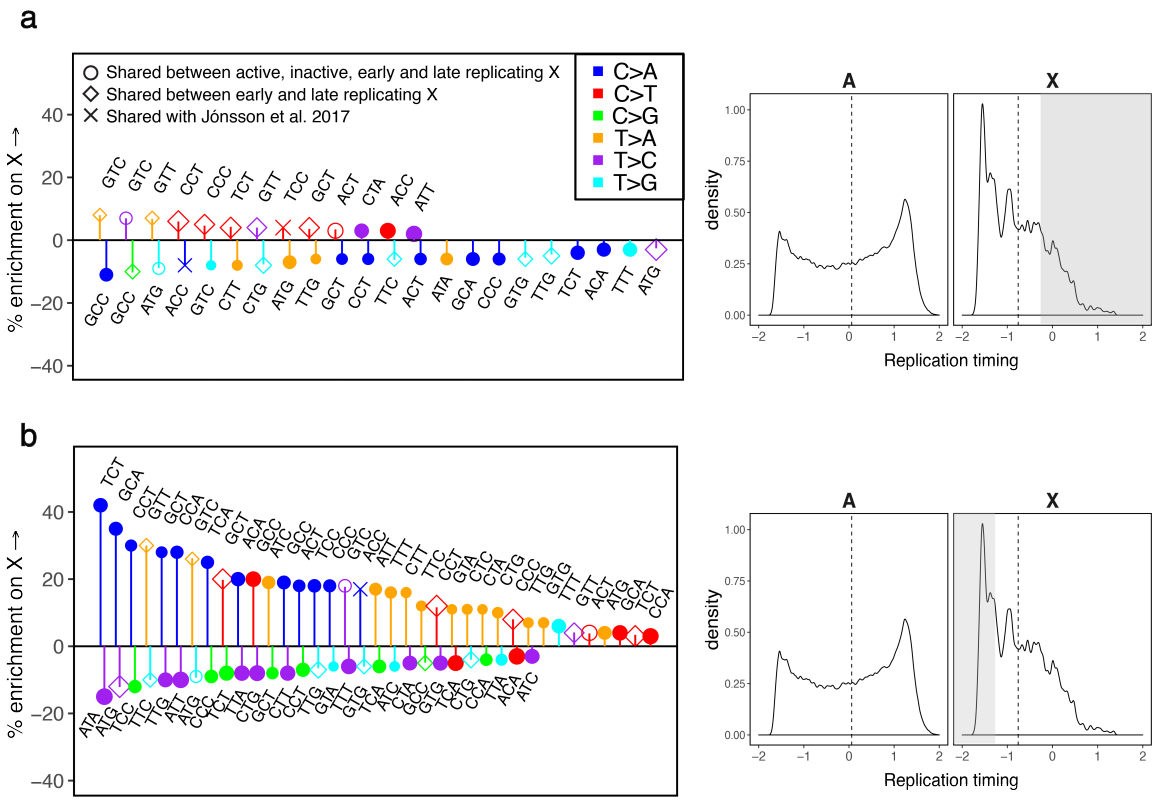

Figure 1.19: Comparison of the mutation spectrum in early and late replicating compartments of the X chromosome versus autosomes. The pseudoautosomal region (PAR) and CpG sites are excluded from this analysis. Only significant differences are shown. Positive and negative effects are ranked separately in order of effect size from left to right. The size of the marker reflects the number of mutations of that type. Hollow circles indicate the three mutation types that are significantly different in both early and late replicating compartments of the X relative to autosomes and also found to be significant differences in both the escaped and inactive compartments of the X relative to autosomes. Crosses denote mutation types reported as significant sex differences by Jónsson et al., 2017. Hollow diamonds represent other mutation types that are significantly different in both early and late replicating compartments of the X relative to autosomes. The distribution of replication timing for the X and autosomes is shown in the right panels, with the mean replication timing on the X and autosomes represented by dashed vertical black lines. Shaded regions indicate the range of replication timing used in the corresponding analysis. (a) Enrichment of mutation types in early replicating regions of the X chromosome (replication timing score $>$ -0.25) relative to autosomes. (b) Enrichment of mutation types in late replicating regions of the X chromosome (replication timing score $<$ -1.25), relative to autosomes.

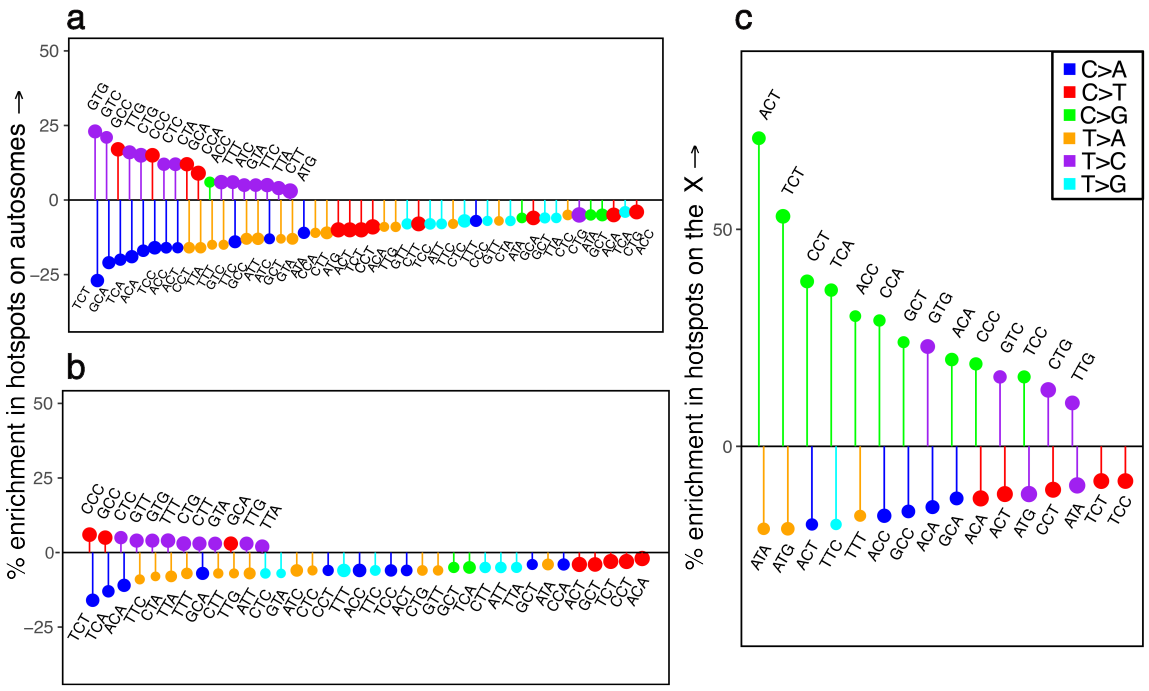

Figure 1.20: Comparison of the mutation spectrum in regions identified as recombination hotspots, relative to autosomal non-hotspots. These analyses exclude regions on autosomes rich in clustered de novo mutations, identified by Jónsson et al., 2017. CpG sites are excluded. Only significant differences are shown. Positive and negative effects are ranked separately in order of effect size from left to right. The size of the circle reflects the number of mutations of that type. (a) The mutation spectrum in hotspots of DMC1-binding measured on autosomes in males, defined as autosomal regions with hotspot intensity > 0. (b) The mutation spectrum in female crossover hotspots, defined as regions with standardized recombination rate > 10. (c) The mutation spectrum in hotspots of DMC1-binding measured on the X chromosome in males, defined as regions outside the PAR with hotspot intensity > 0.

Table 1.1: Sources of whole genome annotation data.

Variants	
gnomAD	gnomAD release 2.0.1 (http://gnomad.broadinstitute.org)
SGDP	SGDP-Lite (http://reichdata.hms.harvard.edu/pub/datasets/sgdp/)
UK10K	ALSPAC and TwinsUK (https://www.uk10k.org/data_access.html)
Genomic features	
Replication timing	LCL (http://mccarrolllab.org/wp-content/uploads/2015/03/Koren-et-al-Table-S2.zip). H1- hESC (wgEncodeFsuRepliChipH1hescWaveSignalRep1); H7-hESC (wgEncodeFsuRepliChipH7esWaveSignalRep1); H9-hESC (wgEncodeFsuRepliChipH9esWaveSignalRep1) from https://genome.ucsc.edu/encode/
CpG islands	http://www.haowulab.org/software/makeCGI/model-based-cpg-islands-hg19.txt
DMC1 ChIP-seq	Human spermatocytes (GEO Accession GSE59836) http://www.ncbi.nlm.nih.gov/geo/
Recombination rate	Standardized female recombination maps (https://www.decode.com/addendum/)
Clustered DNM regions	Supplementary Table 12 (https://www.nature.com/articles/nature24018)
Genic regions	Gencode v19 genes (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz)
Genic regions of X-inactivation and escape	Supplementary Table 13 (https://www.nature.com/articles/nature24265)
Pseudoautosomal region (PAR)	http://genome.ucsc.edu/cgi-bin/hgGateway

Chapter 2

Mutation saturation for fitness effects at human CpG sites

2.1 Abstract

Whole exome sequences are now available for hundreds of thousands of humans. For highly mutable sites, that means we are approaching an important limit, in which datasets are large enough that, in the absence of natural selection, every site will have experienced at least one mutation in the genealogical history of the sample. Indeed, methylated CpG sites that mutate to T at an elevated rate (~10^{-7} per bp per generation) are already very close to meeting that criterion: In a sample of 390,000 individuals, 99% of putatively-neutral, synonymous CpG sites harbor a C/T polymorphism. CpG sites therefore provide a natural mutation saturation experiment for fitness consequences: as we show, at current sample sizes, not seeing a polymorphism is indicative of strong selection against that mutation. We leverage this idea in order

to identify CpG transitions that are too deleterious to segregate in current samples across annotations with similar mutation rates. On that basis, we estimate that whereas ~27% of loss of function mutations are likely to be highly deleterious, only 6% of missense mutations are; however, the proportion increases substantially--up to 21%--depending on the type of functional site in which they occur. As we discuss, in contrast to CpG transitions, mutation types with rates on the order of 10^{-8} or 10^{-9} remain very far from saturation.

2.2 Main text

A long-standing aim of population genetics has been to model the impact of selection on variation within and between species, and to infer fitness effects of mutations from patterns of genetic variation [1, 7, 2, 3, 4]. Because purifying selection depletes deleterious variation over time, it leaves a footprint of conservation in DNA sequences. This signal of reduced genetic variation can then help pinpoint genomic loci of functional importance: for instance, comparisons of sequences across species have been widely used to identify relatively invariant regions, presumably maintained by strong selection over millions of years because of their functional importance [12, 16, 15]. The same general approach is also a useful tool in human genetics, given that genomic sites under purifying selection are enriched for sources of genetic disease and other traits of interest [12, 13, 14, 16, 17]

When using genetic variation within humans to identify sites under selection, the time scale over which mutations may have accumulated is relatively short. The

miniscule mutation rate at a typical site in the genome (~10^{-8} mutations per generation) therefore poses a major difficulty: even sites at which changes have no effect on fitness may not be segregating in available samples. A site could be monomorphic when mutations at that site have no fitness consequences at all or, at the other extreme, when the mutations are embryonically lethal. Consequently, in small samples of a few hundred or even thousands of individuals, with only a small proportion of sites segregating, there is little information on the distribution of fitness effects across the genome. The lack of information about most sites also limits the utility of such samples in distinguishing benign from pathogenic variants in a clinical context.

In part to overcome these limitations, public repositories of human exome sequences have now grown to include data from hundreds of thousands of individuals [119, 13, 14, 120, 121]. In principle then, we should for the first time have at least some direct information about the strength of selection on an unprecedented number of sites in the human genome--particularly at the subset of sites with higher than average mutation rates [13]. To evaluate the merits of this notion, we considered CpG sites methylated in the germline, since these experience mutations much more frequently than any other type of site [23, 20, 21]. We focus on "highly-methylated" CpG sites in exons, defined as those that are methylated >=70% of the time in both testes and ovaries. For these ~1.1 million sites (of 1.8 million total CpG sites in sequenced exons), we calculate a mean haploid C>T mutation rate of 1.17×10^{-7} using de novo mutations (Methods, **Figures 2.5 - 2.6**), an order of magnitude more than the genome average [21].

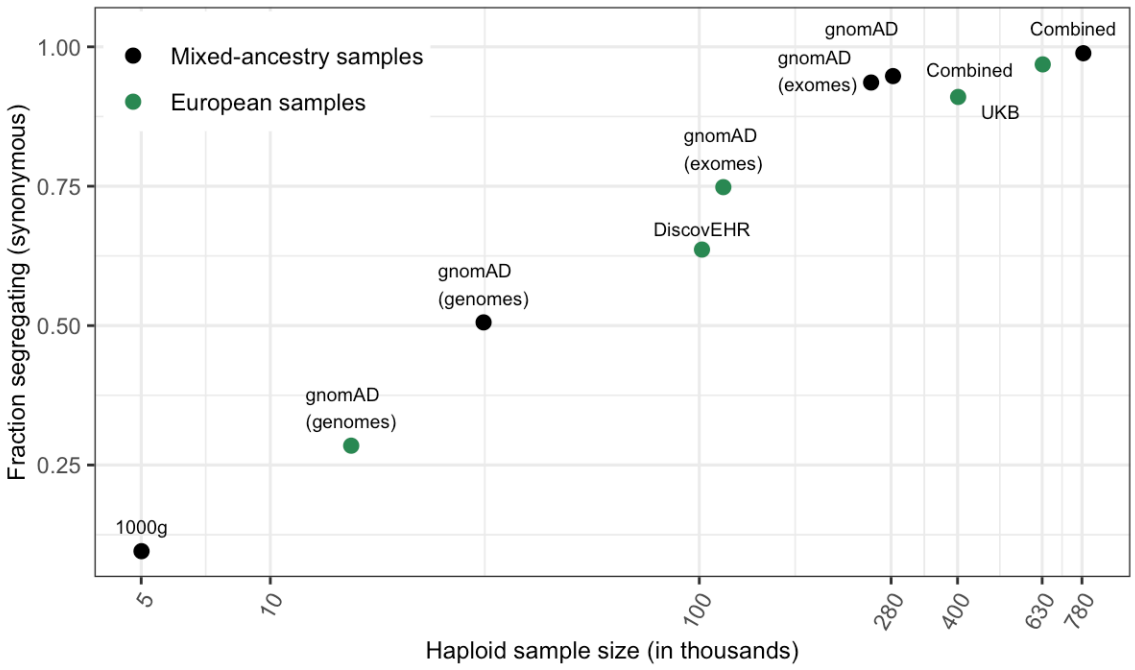

Figure 2.1: Fraction of highly methylated CpG sites that are polymorphic for a transition, for each sample. The combined dataset encompasses three non-overlapping data sources: gnomAD (v2.1), the UK Biobank (UKB), and the DiscovEHR cohort. "European" samples include the "Non-Finnish European" subsets of exome and whole genome datasets in gnomAD, as well as the UK Biobank and DiscovEHR, which have >90% samples labeled as of European ancestry.

Next, we collate polymorphism data made public by gnomAD [13], the UK Biobank [120], and the DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System [119] to ascertain whether a site is segregating in a sample of ~390,000 individuals (Methods). To focus on the subset of genic changes most likely to be neutrally-evolving, we consider the subset of ~350,000 highly-methylated CpG sites at which C>T mutations do not change the amino acid. At these sites, 94.7% of all possible synonymous CpG transitions are observed in the gnomAD data alone, and 98.8% in the combined sample including all three

datasets (**Figure 2.1**).These data indicate that transitions are very close to saturated at highly methylated CpG sites in the sample of 390,000 individuals--in other words, nearly every highly methylated CpG site where a mutation to T is putatively neutral has experienced at least one such mutation in the history of the sample.

If we assume that in the absence of selection, almost every highly methylated CpG site would be segregating a T at current sample sizes, then not seeing a T strongly suggests that it was removed by selection. How strong selection has to be for the site to be monomorphic is unclear, however. Because genetic variation arises from the combined influences of mutation, selection, and random genetic drift over evolutionary time, and because selection strengths and mutation rates vary across genomic loci, isolating the effects of selection requires assumptions about demographic history and the mutation rate at the locus of interest. To examine the relationship between selection and the amount of variation at a locus, we simulate evolution at a single CpG site that undergoes mutations to T at rate 1.2×10^{-7} per generation, under a variant of the widely-used Schiffels-Durbin demographic model for population growth in Europe [9], in which we set the effective population size Ne equal to 10 million for the past 50 generations (Methods). While this model is a vast oversimplification, it mimics the qualitative behavior under neutrality. For instance, in a sample size of 780,000 chromosomes, it suggests that a site experiences at least one mutation in 99% of simulations (**Figure 2.2**). Thus, 99% of neutral sites with this mutation rate are expected to be polymorphic at this sample size, in good agreement with what we observe in data for synonymous CpG transitions (**Figure 2.1**). As expected, the probability that a site is segregating is lower if it is under selection than if it is

neutral; this effect is particularly pronounced for strong selection in smaller sample sizes (**Figure 2.2**).

Although the relationship of selection strengths to clinical pathogenicity is not straight-forward, selection coefficients on the order of 10^{-2} or 10^{-3} are likely to be of relevance to determinations of pathogenicity in clinical settings ([122, 123]; Agarwal, Fuller, Przeworski, in prep.). For instance, many mutations with hs on the order of 10^{-3} may substantially increase liability to a disease; others may on their own be highly deleterious to some individuals that carry them, enough to produce clinically visible effects, but vary substantially in their penetrance. Importantly then, even such highly deleterious mutations are expected to segregate in large samples: in current exome sample sizes, a site with a heterozygote selection coefficient (hs) of 0.5×10^{-3} is almost always observed segregating (**Figure 2.2A**). This follows from the expectation under mutation-selection-drift balance: for example, in a constant population size, a mutation that arises at rate 1.2×10^{-7} per generation and is removed by selection at rate 0.05% per generation is expected in the population at frequency 2.4×10^{-4} on average; in a sample of 780k, the mean number of copies is 187. Thus, even with substantial variation due to genetic drift and sampling error, such a site should almost always be segregating at that sample size. In fact, even a mutation with hs of 5% would quite often be observed. An implication is that, although reference repositories such as gnomAD were partly motivated by the possibility of excluding deleterious variants--with the idea that seeing a variant of unknown function in a reference data set is suggestive that the variant is benign--as samples grow in size, it cannot simply be assumed that deleterious variants are absent from reference datasets. Indeed, with

sufficiently large sample sizes, the only variants ultimately not observed will be those that are embryonic lethal.

Conversely, without information on the mutation rate and sample size, the observation that a site is segregating in a sample only excludes the possibility that the variant is embryonic lethal. In other words, an assessment of the fitness consequences of a site, whether or not segregating, can only be made given a mutation rate and sample size. Moreover, the information content associated with what is segregating or not shifts with sample size. To make this point more concrete, we assume a relatively uninformative log-uniform prior on the selection coefficient s ranging from 10^{-7} to 1 and fix the dominance coefficient h=0.5 (as for semi-dominant mutations, only the compound parameter hs affects allele dynamics; reviewed in [124]). We then estimate the posterior distribution of hs at a site conditional on a mutation rate of 1.2×10^{-7} per site per generation and the demographic model described above, given that the site is monomorphic, segregating with 10 or fewer derived copies of the T allele, or segregating with more than 10 copies (**Figure 2.2**, Methods). It is worth noting here that at current sample sizes, the amount of information per site is limited and thus the choice of prior can be consequential (**Figure 2.7**). Given this sensitivity to priors, and because our demographic model is a clear over-simplification, we do not infer parameters but instead use the model to understand qualitative trends by sample size.

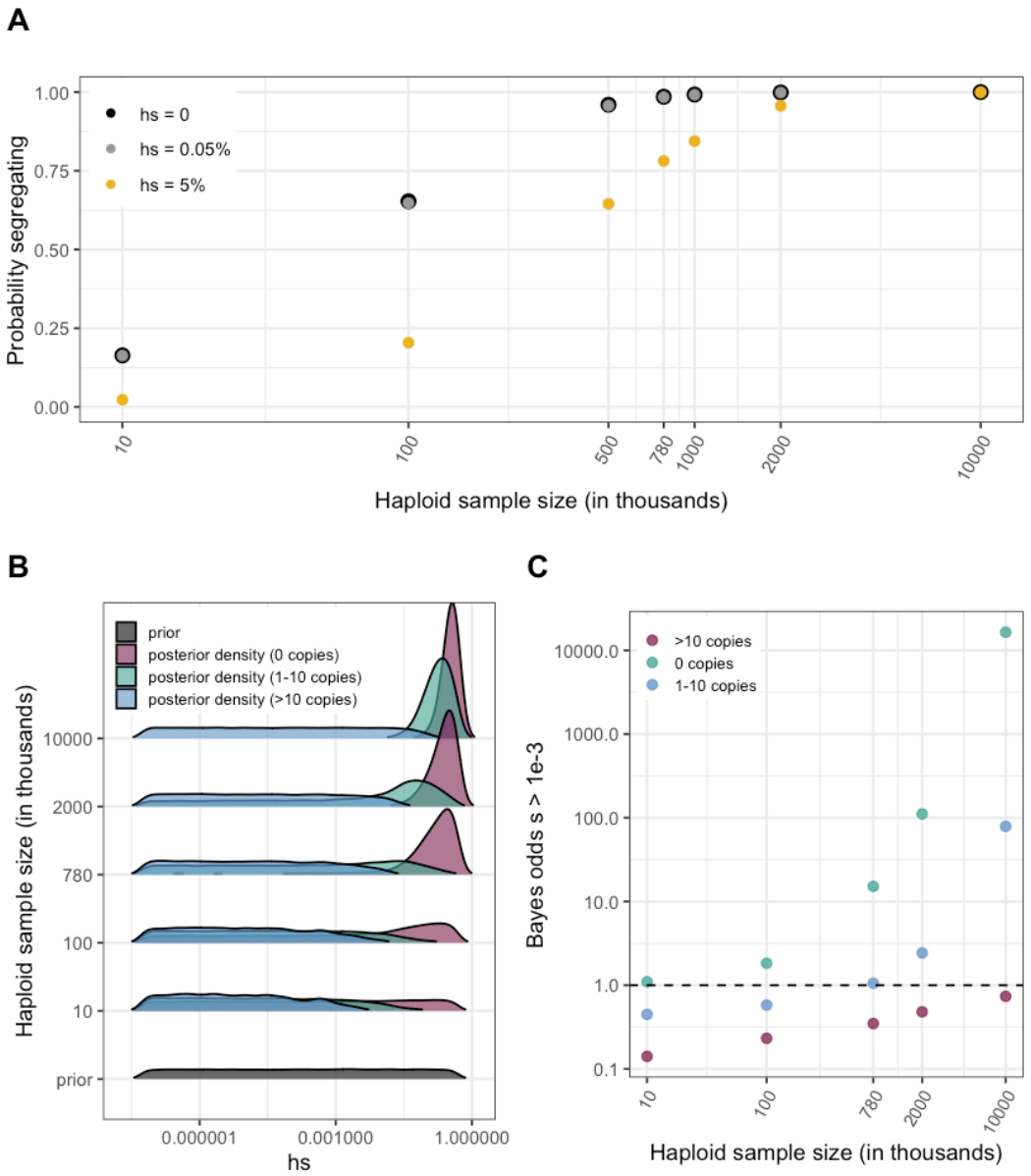

Figure 2.2: **(a)** Probability of a CpG site segregating in simulations, if the mutation has no fitness effects (hs=0) and if it is deleterious (with a heterozygote selection coefficient hs=0.05%) or highly deleterious (with a heterozygote selection coefficient hs=5%, where s is the selection coefficient and h is set to 0.5). **(b)** Prior and Posterior log densities for hs for a mutation observed at 0, 1-10, or >10 copies at various sample sizes. **(c)** Bayes odds (i.e., posterior odds divided by prior odds) of $s > 0.001$ for a mutation observed at 0, 1-10, or >10 copies, at various sample sizes.

At very small sample sizes, most sites are monomorphic, and there is not much information at these sites (i.e., being monomorphic is consistent with both neutrality

and very strong selection). For the small subset of sites that is segregating, the mutation is unlikely to be highly deleterious (and cannot be embryonic lethal): for instance, the posterior odds that a site segregating at 10 or more copies in a sample of 10,000 is under strong selection are 10-fold lower than the prior odds (**Figure 2.2C**). In short, there is almost no information about the overall distribution of fitness effects across sites. In contrast, with larger samples sizes, in which putatively neutral CpG sites reach saturation (~780k; such that that their being invariant by chance is unlikely), the posterior distribution for invariant sites is highly peaked: what is not segregating is likely strongly deleterious. Given our assumptions, the odds that an invariant site in a sample of 780k is under strong selection (set as $hs > 5\times10^{-4}$) are ~10-fold greater than those for sites segregating even at relatively low frequencies.

Somewhat counter-intuitively, with increasing sample sizes, the average polymorphic site becomes consistent with a larger range of selection coefficients, as sites that are under even strong selection start segregating. Thus, the observation that a site is segregating in a large sample is actually less informative about selection than in a small sample, while the opposite is true for invariant sites. Finally, at even larger sample sizes on the order of a few million chromosomes, once neutral and deleterious mutations are potentially present at multiple copies, it becomes possible to distinguish between strongly and weakly selected segregating sites by comparing the frequencies at which they are observed (given an accurate demographic history for the sample). These considerations underscore that interpreting variants of unknown function by reference to repositories such as gnomAD, let alone to disease cohorts that are likely

enriched for deleterious variation (e.g., [125]), become much more complicated in the absence of an underlying model.

In that regard, we note that thus far we have implicitly assumed the same mutation rate at all CpG sites highly methylated in the germline, but this is unlikely to be strictly true, since local sequence context and broader epigenetic features influence local mutation rates [34, 65]. An advantage of conditioning on highly methylated CpG sites is that there is a single known mechanism, i.e., spontaneous deamination of methyl-cytosine, that is believed to be predominantly responsible for their uniquely high mutability [23]. For instance, although regions in and outside exons differ considerably in epigenetic features, replication timing, and the impact of transcription associated damage and repair, there is no appreciable difference in average de novo mutation rates at methylated CpGs inside and outside exons (FET p-value = 0.08, **Figure 2.8**). Thus, while there is likely some residual variation in mutability per site, it is expected to be small relative to the mean mutation rate. At current sample sizes, a small amount of mutation rate variation would be expected to slightly reduce the fraction of synonymous sites saturated, and if not modeled, may be misconstrued as purifying selection at individual sites that are less mutable (**Figure 2.9**).

Given the lack of knowledge of mutation rates at each site and uncertainty about the appropriate demographic model, we cannot infer hs reliably for any given monomorphic CpG site. Instead, we compare the typical fitness effects of CpG transitions across annotations. For this entirely empirical approach, we only need the distribution of mutation rates to be the same across annotation classes and to assume that the distribution of genealogical histories is the same, i.e., that the classes

are subject to comparable effects of linked selection. We therefore verify that the mean de novo mutation rate is the same for synonymous sites as for other annotations (**Figure 2.3**). We also check that the rate at which two DNMs occur at the same site, a summary statistic that reflects the variance in mutation rates, is not significantly different for methylated CpGs inside and outside exons (FET p-value = 0.5; **Figure 2.8**). While the distribution of mutation rates could nonetheless differ somewhat, these observations suggest that the assumption of similar transition rates at highly methylated CpGs across annotations is warranted. The assumption of similar distributions of genealogical history also seems sensible, given that the annotations are closely interdigitated within genic regions; nonetheless, we also check that each comparison yields similar results if the annotations are matched for the predicted effects of linked selection ([126]; see below).

First, we consider the fraction of CpG sites segregating for a transition in each annotation class in a sample of 780k chromosomes, normalized to what is seen at synonymous sites. All categories of missense, loss-of-function, and regulatory variants show a significant depletion in the fraction of segregating sites compared to synonymous variants (**Figure 2.3**). Moreover, the degree of depletion is directly informative about the fraction of sites under selection in each of the annotation classes.

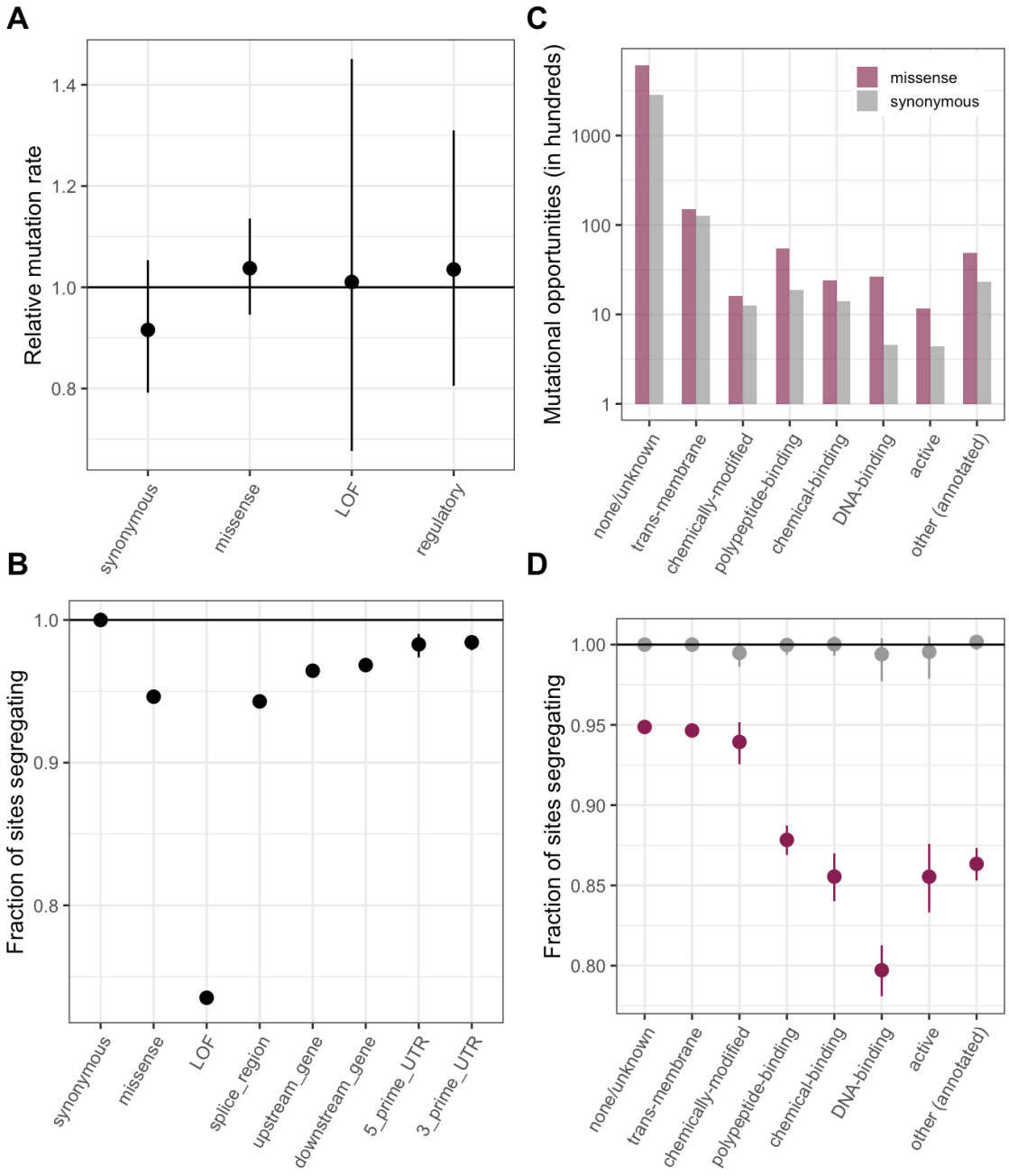

Figure 2.3: **(a)** DNM rates for CpG transitions at highly methylated sites by annotation class, normalized by the total DNM rate in exons **(b)** Fraction of highly methylated CpG sites that are segregating as a C/T polymorphism in an annotation class, relative to the fraction of synonymous sites segregating. Error bars are 95% confidence intervals assuming the number of segregating sites is binomially distributed (Methods). LOF variants are defined as stop-gained and splice donor/acceptor variants that do not fall near the end of the transcript, and meet the other criteria to be classified as "high-confidence" loss-of-function in the gnomAD data. **(c)** The number of opportunities for synonymous and missense changes involving highly methylated CpG transitions by the type of functional protein site. **(d)** The proportion of synonymous and missense segregating C/T polymorphisms in different classes of functional sites. All annotations are obtained using the canonical transcripts of protein coding genes (see Methods).

Notably, these data suggest that there are ~27% fewer loss-of-function variants than would be expected under neutrality. With our model and our relatively uninformative prior, an invariant site in this sample of 390,000 individuals is very likely to have an $hs > 0.5\times$1e-3 (posterior odds ~11.5 to 1, **Figure 2.2**). If we assume that all true LOF mutations in a gene (after filtering for those at the end of transcripts, for instance) are identical in their fitness cost, as is standard (e.g., [122, 13]), then CpG LOF mutations should be informative about the general distribution of consequences for LOF mutations. Supporting the assumption that true LOF mutations within a gene are exchangeable, when we compare the set of CpG sites at which mutations lead to possible protein-truncation in the first and second half of transcripts respectively, approximately the same number are under strong selection in both subsets (**Figure 2.10**; FET p-value ~0.09).

In contrast, only 6% of missense variants and splice region variants appear to be under the same degree of selection (whether or not we match for the effects of linked selection; see **Figure 2.11**). While LOF and missense annotation classes are most commonly used in determinations of variant pathogenicity, any two sets of highly methylated CpGs with similarly-distributed mutation rates can be ranked. We therefore consider the fitness effects of missense mutations, stratifying them by the type of functional site in which they occur. Strikingly, for the subset of sites at which missense mutations may disrupt or alter binding, particularly DNA-binding, ~21% of variants are likely to be under strong selection, in contrast, say, to missense changes within trans-membrane regions (**Figures 2.3, 2.12**, 2.11). Thus, a more fine-grained classification of missense mutations reveals strong selection acting on a

subset--on par with what is seen for LOF mutations.

Another common approach is to group sites by conservation scores or measures derived from conservation scores such as CADD [127]. As expected, the fraction of sites segregating decreases with increasing CADD scores, with 17% of highly methylated CpG sites monomorphic for the top decile of CADD scores (**Figure 2.13**) We note that for these CpG sites, mean mutation rates are similar across CADD deciles, as expected if DNMs are rarely embryonic lethal, and if there is limited variation within this set. Importantly, however, this is no longer the case when considering all CpG sites in exons: for this larger set of sites, the mean de novo mutation rate varies across deciles of CADD, potentially because these scores, while meant to isolate the effects of selection, nonetheless confound mutation rates and conservation to some extent (**Figure 2.13**). We therefore caution that included among relatively high CADD scores in the genome are likely some sites that are not in fact unusually constrained, but have a lower mutation rate.

Given that current exome samples are informative about selection at highly methylated CpGs, a natural question is to ask to what extent there is also information for less mutable types, with mutation rates on the order of 10^{-8} or 10^{-9}. That synonymous CpG sites are close to saturation when they experience mutations to T at a rate of of 1.17×10^{-7} per generation implies that, on average, the total branch length of the genealogy relating the 390K individuals is at least $0.85\text{x}10^{7}$ generations (i.e., at a site, the sum of the branch lengths of the genealogy is on average greater than 1 divided by 1.17×10^{-7}). Assuming the average length of the genealogy is comparable or shorter for other types of sites (**Figure 2.14**), we should therefore expect

that most sites with much lower mutation rates are not segregating. Indeed, for sites with mutation rate on the order of 10^{-9}, which is the case for the vast majority of non-CpGs, the fraction of possible synonymous T>A variants observed is ~4% in the sample of 780K chromosomes, compared to 99% for C>T variants at highly methylated CpGs and ~30% for all other C>T variants (**Figure 2.4**). Thus, at current sample sizes, there is not much information about selection for these mutations. One implication is that the genealogical history of this sample is over 0.85×10^7 generations long, but substantially shorter than 10^8 generations.

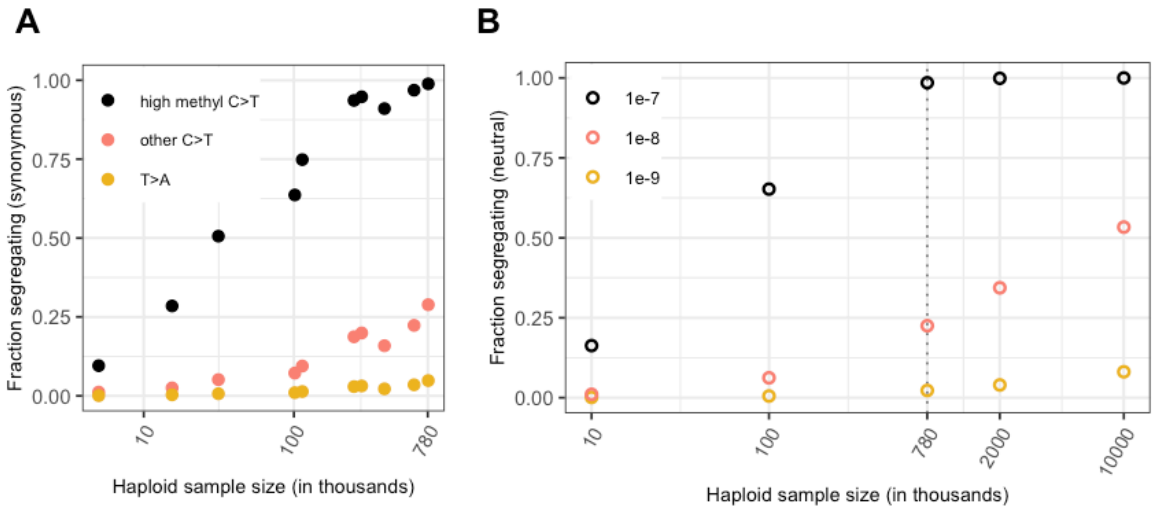

Figure 2.4: **(a)** Fraction of possible synonymous C>T mutations with high levels of methylation in the germline, other C>T mutations, and the fraction of possible synonymous T>A mutations that are observed in a sample of given size. **(b)** Probability of a polymorphism in simulations, assuming neutrality, a specific demographic model and a given mutation rate (see Methods).

Since the length of the genealogy does not increase linearly with the number of samples, as is well appreciated from coalescent theory [5], saturation may not even be achievable in extremely large samples on the order of tens of millions for such mutation types. Under our choice of demographic model and assuming neutrality, a

sample of 780,000 chromosomes has a genealogy spanning an average of 40 million generations. Increasing the number of samples by a factor of 12 only increases tree length ~3.3x (**Figure 2.4**, **Figure 2.15**); thus, a site that mutates at rate 10^{-9} per generation is expected to have experienced ~0.04 mutations in the genealogical history of a sample of ~1 million, and 0.1 mutations in a sample of 10 million. For most types, therefore, mutation saturation is far off, and information on selection is limited to the few sites that are segregating in current samples (**Figure 2.4**, **Figure 2.16**).

Quantitative predictions of our model are unreliable, given uncertainty about the demographic history and in particular the recent effective population size in humans (**Figure 2.15**) Moreover, for simplicity, we model one or at most two populations, when mixed ancestry samples have slightly longer genealogical histories (**Figure 2.15**). Thus, as samples diversify by ancestry, we will capture slightly more variation than if we only include individuals from similar ancestries. We also note that for the very large sample sizes considered here, the multiple merger coalescent is the more appropriate model [128]. Nonetheless, the qualitative statement that less mutable types remain very far from saturation will hold in the foreseeable future.

In this light then, it is of interest to ask whether the distribution of fitness effects at CpG sites provides a reasonably good approximation for what we might expect to see at other types of sites in the exome. In comparison to other types of sites, CpGs are somewhat enriched for synonymous sites and depleted for missense sites compared to other mutation types and the distribution of CADD scores slightly (though given the amount of data, highly significantly) skewed towards lower values (KS test p-value $<< 10^{-5}$; **Figure 2.17**), as perhaps expected from the behavior of CADD scores in the

presence of mutation rate variation. These differences are subtle, however, suggesting that the distribution of fitness effects at highly mutable CpG sites is a useful proxy for what to expect for the rest of the exome.

Previous work on inferring this distribution in humans and other organisms relied on very small sample sizes and thus was forced to assume a specific and arbitrary parametric form, inferred with very little or no information about moderately or strongly selected sites [129, 130, 131]. By leveraging the vast exomic data sets now assembled for humans, it is possible to access these questions much more directly, in order to both identify specific sites likely to be under strong selection and quantify the fraction of highly deleterious mutations across annotations. There is still not enough information to precisely estimate the strength of selection at individual sites in the genome, however. For methylated CpG sites, which have a mutation rate of 10^{-7}, this should become possible with samples of a few million individuals. Determining the distribution of selection coefficients in the human genome is therefore within reach, at least through the lens of CpG sites and provided a good characterization of mutation rate variation within this class. More generally, better predictions of mutation rates from de novo mutation data or models that rely on our increasing understanding of mutational mechanisms should enable inferences about fitness effects in other mutational contexts.

2.3 Acknowledgements

We thank Hakhamanesh Mostafavi, Jonathan Pritchard, Magnus Nordborg, as well as Arbel Harpak, Zach Fuller, Guy Sella and other members of the Andolfatto, Przeworski and Sella labs for helpful discussions. This work was supported by NIH grants GM121372 and GM122975 to MP.

2.4 Materials and Methods

2.4.1 Processing de novo mutation data

We obtained ~190,000 published de novo mutations in a sample of 2976 parent-offspring trios that were whole genome sequenced [26]. To date, this is the largest publicly available set of trios that, to our knowledge, have not been sampled on the basis of a disease phenotype. Unless otherwise specified, we use these DNMs to calculate mutation rates, as described in later sections. We converted hg38 coordinates to hg19 coordinates using UCSC Liftover. We excluded indels, and all DNMs that occur outside the ~2.8 billion sites covered by gnomAD v2.1.1 whole genome sequences. We obtained the immediately adjacent 5' and 3' bases at each position from the hg19 reference genome, so that we had each de novo mutation within its trinucleotide context; we used this information to identify CpG sites. Where such data were available (for 89% of CpG de novo mutations), we also annotated each site with its methylation status in testes and ovaries (see **Table 2.1**).

2.4.2 Processing polymorphism data

We downloaded publicly available polymorphism data from gnomAD [13], the UK Biobank [120], the DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System [119], and 1000 Genomes Phase 3 [71]. Where needed, we lifted over coordinates to the hg19 reference assembly using the UCSC LiftOver tool. Salient characteristics of these samples are summarized below.

For the gnomAD data, we obtained the allele frequency for each variant in the full exome and genome samples, as well as their Non-Finnish European ("NFE") subsets from the VCF files (in hg19 coordinates) provided. For each sample, we obtained the set of segregating sites (i.e., the set of variants that pass gnomAD quality filters and have an allele frequency > 0 in the sample). For the 1000 Genomes Phase-3 data, we similarly obtained the set of variant positions. Note that the 1000 Genomes samples are also contained within the gnomAD sample. For the DiscovEHR sample, allele frequencies are available where MAF > 0.001 (and set equal to 0.001 for lower values>0); this information allows us to determine the set of sites segregating in this sample, but we do not have access to any other information about individual variants.

Additional processing was required for the UK Biobank exome sequencing data. We downloaded the population-level plink files with exome-wide genotype information for ~200,000 individuals. We excluded exome samples that did not pass variant or sample quality control criteria in the previously released genotyping array data. Specifically, we excluded samples that have a discrepancy between reported sex and inferred sex from genotype data, a large number of close relatives in the database, or

Dataset	Sequenced regions	Individuals	Variants	Populations sampled
gnomAD v2.1.1	Exomes	125,748	15 million	mixture
gnomAD v2.1.1	Genomes	15,708	230 million	mixture
UK Biobank	Exomes	200,643	16 million	~93% European ancestry
DiscovEHR	Exomes	50,726	8 million	~98% European ancestry
1000 genomes Phase 3 (also included in gnomAD)	Genomes	2,504	84 million	mixture

are outliers based on heterozygosity and missing rate, as detailed in Bycroft et al., 2018 [132]. Finally, we excluded individuals who withdrew from the UK Biobank by the end of 2020. This left us with 199,930 exome samples that overlap with the high-quality subset of the genotyped samples. We additionally limited our analysis to the list of ~40 million exonic sites with an average of 20x sequence coverage provided by UK Biobank, and for which variants met the QC criteria described in Szustakowski et al., 2020. We transformed the processed plink files into the standard variant call format, polarized variants to the hg38 reference assembly, and obtained the frequency of the non-reference allele in the sample. We then lifted over the coordinates from hg38 to hg19 using the UCSC LiftOver tool. We excluded the few positions where the reference alleles were mismatched or swapped between the two assemblies.

2.4.3 Identifying and annotating mutational opportunities in the exome

For all possible mutational opportunities in sequenced exons, we collated a variety of functional annotations. To this end, we first generated a list of all possible SNV mutational opportunities in the exome. We obtained the list of sites that fall in exons or within 50bp of exons in Gencode v19 genes and that are among the ~2.8 billion sites covered by gnomAD v2.1.1 whole genome sequences. For each position, we extracted the reference allele from the hg19 assembly and generated the three possible single-nucleotide derived alleles. We also obtained the immediately adjacent 5' and 3' bases at each position from the hg19 reference genome, so that we had each mutational opportunity within its trinucleotide context; we used this information to identify CpG sites. Where such data were available, we also annotated each site with its methylation status in testes and ovaries.

To identify sites at which variants or de novo mutations could be confidently assayed by whole-exome sequencing methods, we obtained regions targeted in whole exome sequencing from gnomAD and the UK Biobank. We limited our analysis to sites that were covered at 20x or more in the exome sequencing subsets of both gnomAD and UK Biobank (that lifted over correctly to the hg19 assembly), which we refer to as "accessible sites".

We then annotated the ~90 million mutational opportunities (at 30 million sites) with CADD scores and variant consequences using Variant effect predictor (v87, Gencode V19) and the hg19 LOFTEE tool [13] to flag high-confidence ("HC") loss-

of-function variants. For loss-of-function variants, we also noted their location in the gene by exon number (e.g., in exon 10 of 12 exons in the gene). We used a database of curated protein features derived from Refseq proteins [133] to annotate all sites in protein coding genes that were associated with a particular type of functional activity (detailed functional annotations were available for 62,387 of 1.1 million highly methylated CpG sites). At each site, we used either the primary "site-type" annotation, or when that was missing or listed as "other", we extracted the annotation from the more detailed "notes" field where this information was provided.

Because there are multiple transcripts for each variant, we limited our analysis to the "canonical" protein-coding transcript for each gene provided by Gencode to obtain a single annotation for each variant. For 10-20% of variants this approach still yielded multiple possible consequences per variant, for instance, where there are multiple canonical transcripts due to overlapping genes. For these cases, we assigned one of the "canonical" transcripts to the variant at random, to avoid making assumptions about their relative importance. Further overlaps within the same gene, e.g. a missense variant that is also a splice variant in the same transcript, or a DNA-binding site that also undergoes a particular post-translational modification were resolved in the same manner.

As an alternative approach, we obtained the worst consequence in all protein coding transcripts for each variant, using the ranks of variant consequences by severity provided by Ensembl (see **Table 2.1**). In the absence of systematic ranking criteria for the protein function annotations we used the following order: sites that were designated as having catalytic activity ("active" sites) were given highest priority in

overlaps, followed by DNA-binding sites, followed by other types of binding (to metal, polypeptides, ions), and finally by sites that are known to undergo post-translational or other regulatory modifications, and trans-membrane sites. Thus, a transmembrane site with regulatory activity is classified as a regulatory site, while a regulatory site with DNA-binding activity is classified as DNA-binding. Using these alternate criteria to group sites does not affect our conclusions (**Figure 2.12**).

All sources of annotation data are listed in (**Table 2.1**).

2.4.4 Comparing fitness effects across sets of mutational opportunities

To assess whether the set of 1.1 million C>T mutational opportunities at highly methylated CpG sites are systematically different from the other ~90 million exonic mutational opportunities in their potential fitness effects, we compared the distribution of CADD scores in the two groups using a Kolmogorov-Smirnov test. We note that this comparison is likely to be somewhat confounded by differences in mutation rates, given our finding that CADD scores do not perfectly isolate the effects of selection from those of variability in mutation rates (**Figure 2.13**). Since the mutation rate for highly methylated CpG sites is higher than for other types, they may appear somewhat less constrained than they actually are.

We further compared the fraction of C>T mutational opportunities at highly methylated CpGs in an annotation class vs. the fraction of other mutational opportunities in that class. We used a Fisher exact test (with a Bonferroni correction

for four tests) to determine whether the two sets of mutational opportunities were differently distributed across synonymous, missense, regulatory, and LOF variant classes.

2.4.5 Obtaining mean de novo mutation rates by mutation type and annotation

We counted the total number of de novo mutations in sequenced exons (~91 million mutational opportunities) for 8 classes of mutations: two transitions and a transversion each at C and T sites, transitions at CpG sites with relatively low levels of methylation (i.e., methylated $< 70\%$ of the time in testes and ovaries, measured by bisulfite sequencing), and transitions at CpG sites with high levels of methylation (>=70% of the time). To obtain the mutation rate per site per generation, we divided the counts by the haploid sample size (2×2976 individuals) and the number of mutational opportunities of each type. We report 95% confidence intervals assuming a Poisson distribution for mutation counts. The rates (**Figure 2.5**) are roughly consistent with rates predicted by the gnomAD mutation model [13], and similar to previous estimates [22, 21]. We note that an implicit assumption is that the distribution of parental ages in the trio data is representative of the parental ages over the evolutionary history of exome samples, since the mutation rate increases with paternal and maternal ages.

To evaluate the impact of methylation status on the mutation rate at CpG sites, we obtained the mean mutation rate for C>T mutations at CpG sites in each methylation

bin as described above, separately for methylation levels in ovaries and testes. While there is a limited amount of data, especially for some low-methylation bins, our choice of cutoff for "highly methylated" seems sensible (**Figure 2.6**).).

We then calculated the mean mutation rate for highly methylated CpG transitions, for different compartments in the genome, namely in (a) exons and non-exons, (b) four variant consequence categories: synonymous, missense, regulatory, and LOF variants, (c) CADD score deciles, and (d) in exons that constitute the first half vs the second half of genes. In each case, we obtained the total number of de novo mutations and the Poisson 95% confidence interval around mutation counts in each group, and divided by the number of mutational opportunities in the group. We tested if the mutation counts in each compartment were different from the expected counts using the highly methylated CpG transition rate averaged across all compartments using a Poisson test (we implicitly assume a small sampling error for the highly methylated CpG transition rate averaged across all compartments).

2.4.6 Variance in mutation rate at highly methylated CpGs

Although current samples of DNM data are large enough to compare the mean mutation rate at methylated CpGs across the annotation classes we are interested in, there is not enough data to directly compare variances in mutation rates. To learn how much variation in mutation rates at highly methylated CpGs may exist across annotations, we therefore have to rely on a broader set of regions e.g., those that fall inside and outside exons. Exonic and non-exonic regions differ considerably in epige-

netic features, replication timing, and the impact of transcription associated damage and repair; yet, there is no discernable difference in average de novo mutation rates at methylated CpGs inside and outside sequenced exons (FET p-value = 0.08, **Figure 2.8**). We also compared the number of double and single de novo hits in exons and non-exons using a Fisher exact test (p-value = 0.5, **Figure 2.8**). Since the number of double hits reflects the variance in mutation rates across sites, these results lend some support to our assumption of the same distribution of transition rates at highly methylated CpGs across annotations.

2.4.7 Calculating the fraction of sites segregating by annotation

For each highly-methylated CpG site in the exome, there are three mutational opportunities (C>A, C>G, C>T); we focus only on the opportunities for C>T mutations. For each highly methylated CpG site then, we noted whether or not it was segregating, or in other words if it had a C>T variant in samples of individuals from gnomAD [13], the UK Biobank [120], the DiscovEHR collaboration between the Regeneron Genetics Center and Geisinger Health System [119], and 1000 Genomes Phase 3 [71], processed as described above, or a combined sample of 390,000 non-overlapping individuals.

Within the set of methylated CpG sites where C>T mutations are synonymous, we calculated the fraction segregating in each sample of interest. Similarly, for different subsets of methylated CpGs, namely those in (a) four variant consequence categories: synonymous, missense, regulatory, and LOF variants, (c) CADD score deciles, (d)

functional site categories (e.g., trans-membrane vs catalytic sites in proteins), and (e) the first half vs the second half of genes, we calculated the fraction segregating in the combined sample of 390,000 individuals. We normalized the fraction of sites segregating in each annotation by the fraction of synonymous sites segregating in the sample.

We verified that the differences in the fraction of sites segregating across annotations are not due to different effects of linked selection by annotation. To do so, we calculated the fraction of sites segregating with sites in different annotations matched for B-statistics [126]; we obtained very similar results with this approach (**Figure 2.11**).

We assume that conditional on the number of mutational opportunities and a fixed probability of segregating for each site in a compartment, the number of sites segregating is binomially distributed, and obtained 95% confidence intervals on that basis. We tested if the probability of segregating in each compartment was different from the probability of segregating at putatively neutral (here, synonymous) sites using a Binomial test.

We also calculated the fraction of other types of synonymous sites segregating in each sample size of interest (specifically, for T>A variants, and C>Ts not at highly methylated CpG sites).

2.4.8 Forward Simulations

We used a forward simulation framework initially described in Simons et al. (2014) [134], modified in Fuller et al. (2019), and also described in Agarwal, Fuller, and Przeworski (in prep). Briefly, we modeled evolution at a single non-recombining bi-allelic site, which undergoes mutations each generation at rate $2N_e u$ in a panmictic diploid population of effective population size N_e. Each generation is formed by Wright-Fisher sampling with selection, where fitness is reduced by hs in heterozygotes and s in homozygotes for the T allele. We fixed the dominance coefficient h as 0.5, as only the compound parameter hs is important to the dynamics of dominant alleles (reviewed in Fuller et al. 2019), and we choose one value of s for each simulation. Given a mutation rate and a demographic model that specifies N_e in each generation, we simulated the evolution of this locus forward in time to determine whether the site is segregating in a sample of size n at present.

We used $u = 1.2\times10^{-7}$ to model CpG>TpG mutation at a highly methylated CpG site. The simulation framework allows for recurrent mutations, which are expected to arise often at this mutation rate. We also allowed for TpG>CpG back mutations at the rate of 5×10^{-9} (calculated from de novo mutation data, as for CpG>TpG mutations).

We used the Schiffels-Durbin model for population size changes in Europe over the past ~55,000 generations, preceded by a ~$10N_e$ generation burn-in period of neutral evolution at an initial population size N_e of 14,448 (following Simons et al. 2014). In the last generation, i.e., at present, we sample n individuals from the simulated

population, to match the size of the sample of interest.

We calculated the probability that a site with the fixed mutation rate u above is segregating for a given value of h and s (with s=0 under neutrality) as the proportion of simulations with those parameters in which the site is segregating for different sample sizes at present. We also obtained the probability that a neutral site is segregating if the mutation rate is not fixed but rather drawn from a lognormal distribution with mean 1.2×10^{-7}, and for different degrees of variance, as described in Harpak et al. 2016 [24].

In comparing the output of these simulations to data, we consider two scenarios where we may either undercount or overcount segregating CpG sites in the data relative to the simulations. First, because we condition on the human reference allele being CpG in data, we do not count sites where the CpG is the ancestral but not the reference allele. To check how often this is expected to occur, we mimic this scenario in simulations, sampling a single chromosome at the end of the simulation as reference. The proportion of simulations in which CpG is the ancestral but not the reference allele is ~0.1%. The second case is that for a subset of the CpG>TpG variants observed at present, the CpG mutation is not ancestral, yet we do not include that case in simulations. To mimic this scenario in our simulations, we simulate a site that starts as TpG (with a mutation rate of 5×10^{-9} to CpG, and a back mutation rate ~1.2×10^{-7} to TpG) forward in time. Then, as above, we draw a single chromosome from the sample at the end of the simulation and set it as the reference. We obtain the proportion of simulations in which the C allele is the reference, starting from a TpG background. Reassuringly, this occurs in only 0.0014% of simulations. We

note that there is a third scenario, in which ApG or GpG sites is ancestral and a C/T polymorphism is found in the sample at present as a result of two mutations, one to T and one to C. Given the various mutation rates involved (all less than 5 $\times 10^{-9}$), this double mutation case will be even less likely than the one in which TpG was ancestral. These rare scenarios should not have any substantive effect on our comparison of data to simulations, particularly when we only use such comparisons to examine qualitative trends.

2.4.9 Inferring selection in simulations

In lieu of calculating the probability that a site segregates for a fixed value of s, we can propose s from a prior distribution and infer the posterior distribution of hs for a site with a T allele at 0 copies using a simple Approximate Bayesian Computation (ABC) approach. Specifically, we propose s such that $log_{10}(s) \sim U(-7, 0)$; we simulate expected T allele counts under our model for 10 million proposals from the prior. We accept the subset of the proposed values of s where simulations yield 0 copies of the T allele in the sample at present; this set of s values is a sample from the posterior distribution of s given that the site is monomorphic. We calculated the Bayes odds of s $>10^{-3}$ as the ratio of the posterior odds of s $>10^{-3}$ and the prior odds of s $>10^{-3}$:

$$\frac{p(hs > 5 \times 10^{-4} \mid \ T{=}0\ ) / p(hs \leq 5 \times 10^{-4} \mid \ T{=}0\ )}{p(hs > 5 \times 10^{-4)} / p(hs \leq 5 \times 10^{-4})}$$

We similarly obtain posterior distributions of hs (h fixed at 0.5) for sites that

are segregating at 0, 1-10 copies, or >10 copies, in samples of different sizes, and for three different choices of priors on s, namely: $s \sim Beta(\alpha = 0.001, \beta = 0.1)$; $log(s) \sim N(-6, 2)$; and $N_e s \sim Gamma(k = 0.23, \theta = 425/0.23)$, with N_e=10,000, based on the parameters inferred in Eyre-Walker et al. 2006. These are shown in **Figure 2.7**).

2.4.10 Coalescent Simulations to obtain the length of genealogy of large samples

We simulated the genealogy of a sample of varying sizes using msprime [135] under different demographic histories, modifying the standard Schiffels-Durbin models as follows:

(a) Demographic history for a sample of Utah residents with Northern and Western European ancestry (CEU) over 55,000 generations, modified from from Schiffels and Durbin 2014, with a recent Ne of 10 million for the past 50 generations, described above.

(b) CEU demographic history for 55,000 generations with a recent Ne of 100 million for the past 50 generations.

(c) CEU demographic history for 55,000 generations with 5% exponential growth for the past 200 generations.

(d) Demographic history for a sample of Yoruba (YRI) ancestry from Schiffels and Durbin 2014, modified with a recent Ne of 10 million for the last 50 generations.

(e) A structured sample from two populations that derived from an ancestral population with YRI demographic history 2,000 generations ago, with YRI and CEU demographic histories respectively since, and a recent Ne of 10 million for the last 50 generations in each.

The code for implementing these different demographic models in msprime is available as part of the Supplementary Materials. In each case, we recorded the mean genealogy length over 20 iterations.

2.5 Supplementary Tables and Figures

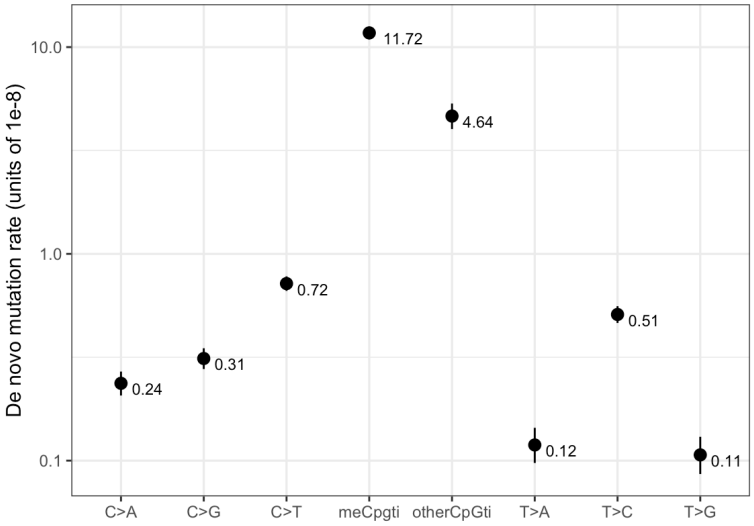

Figure 2.5: Exonic de novo mutation rates in a sample of 2976 parent-offspring trios, by mutation type. Error bars reflect the 95% Poisson confidence interval around mutation counts for each type.

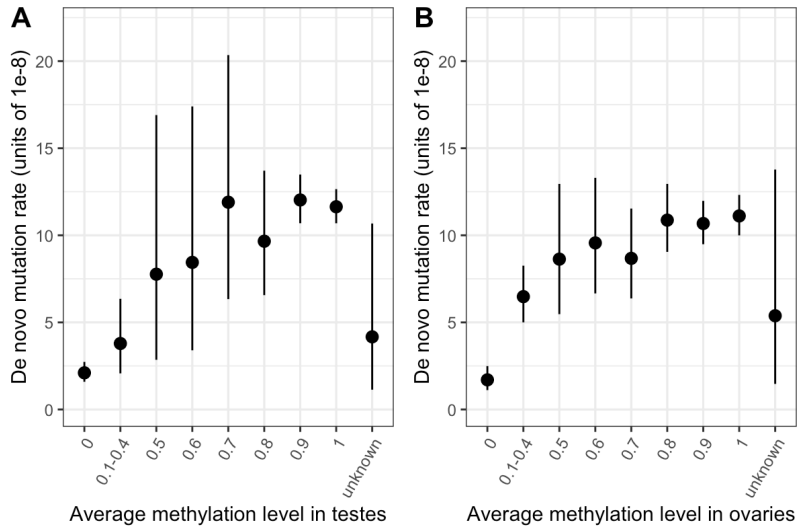

Figure 2.6: De novo mutation rates in exons in a sample of 2976 parent-offspring trios, by average methylation levels in testes and ovarian tissue. Error bars reflect the 95% Poisson confidence interval around mutation counts in each group (the minimum number of DNMs in each bin is 5).

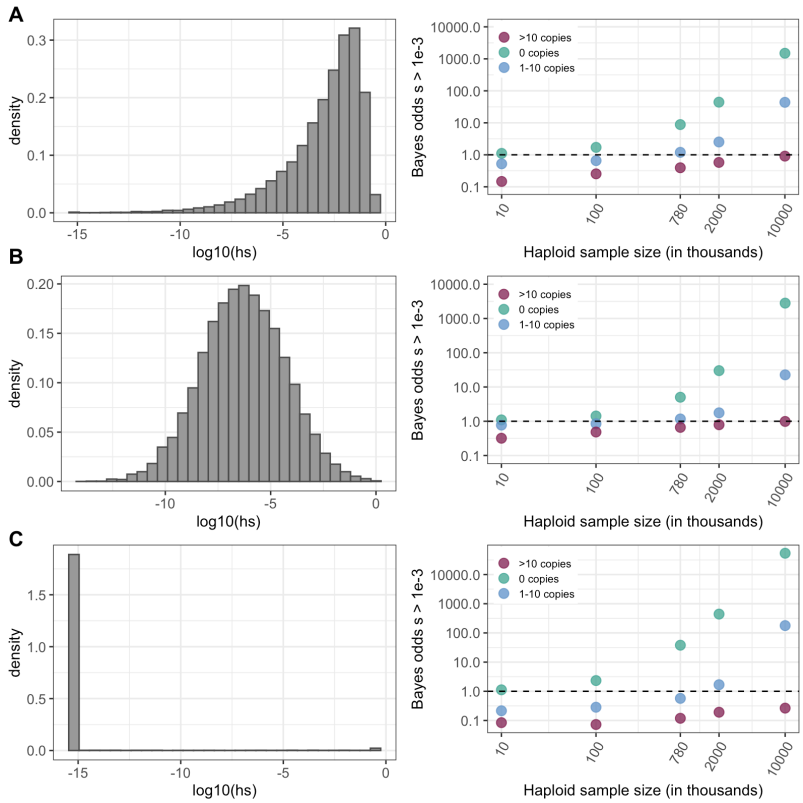

Figure 2.7: Prior on hs (left column) and Bayes odds that s > 0.001 given that a mutation at a site is observed at 0, 1-10, or >10 copies, for various sample sizes (right column). The odds are calculated using 10,000 draws from the prior and posterior distributions. h is fixed at 0.5 (a) $s \sim Beta(\alpha = 0.001, \beta = 0.1)$. Values below 10^{-10} are binned as 10^{-10}. (b) $log(s) \sim N(-6, 2)$. (c) $N_e s \sim Gamma(k = 0.23, \theta = 425/0.23)$, with N_e=10,000, based on the parameters inferred in Eyre-Walker et al. 2006.

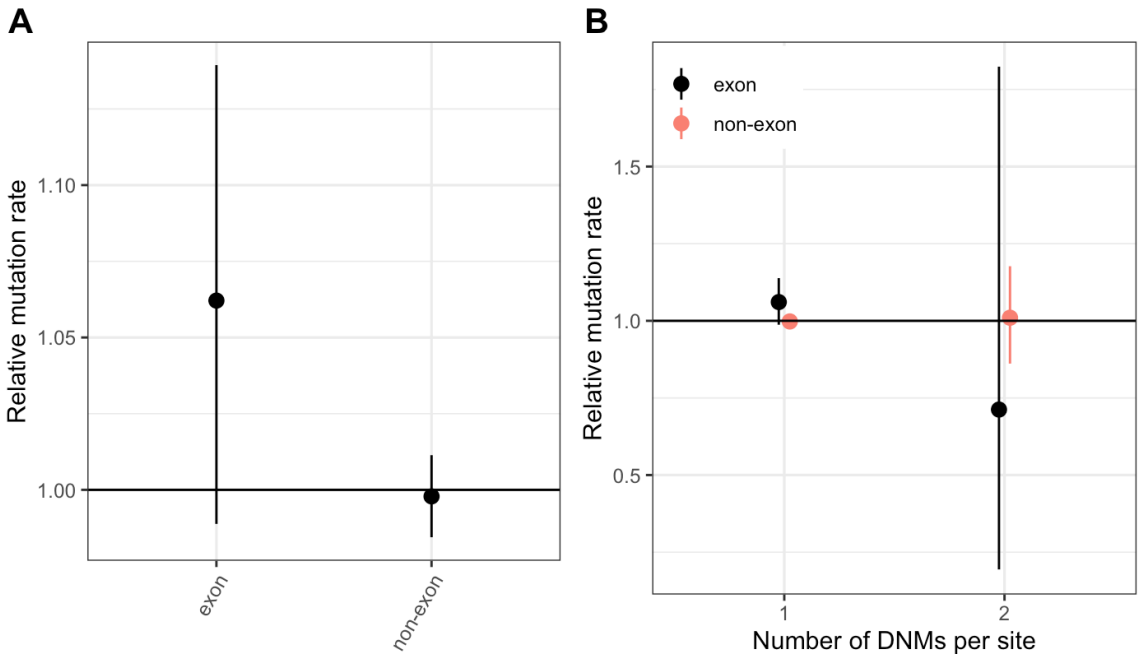

Figure 2.8: Comparing the distribution of mutation rates in non-exons and exons (a) DNM rates for CpG transitions at highly methylated sites in exons vs. non-exons, normalized by the total DNM rate in the genome, with 95% Poisson confidence intervals. (b) The rate of single hits (one DNM at a site) and double hits (two DNMs at a site) in exons vs non-exons, normalized to the average rate of single and double hits in the genome, with 95% Poisson confidence intervals.

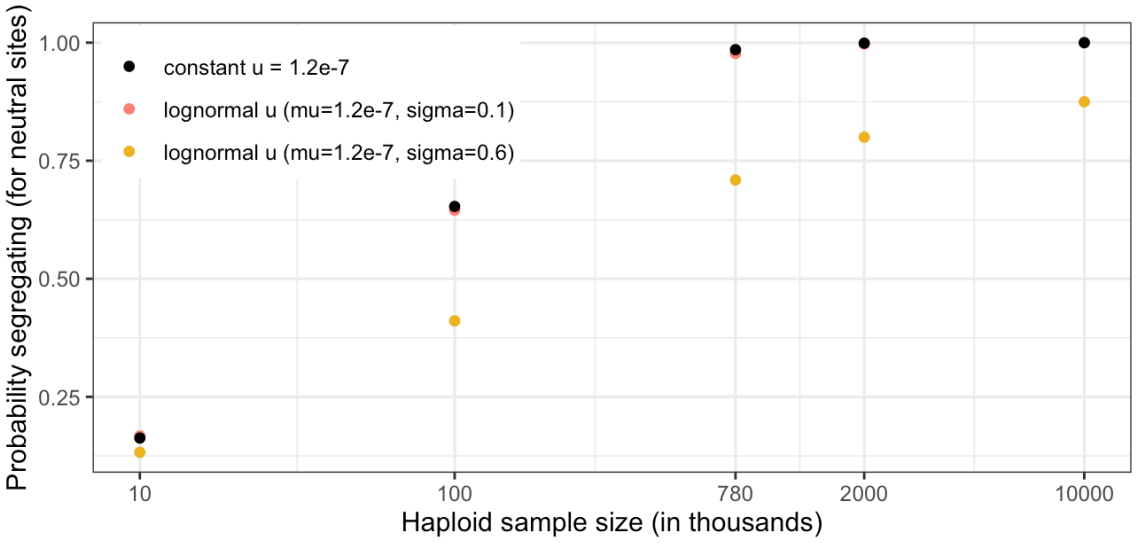

Figure 2.9: The effect of mutation rate variation on the probability that a site is observed segregating under neutrality at different sample sizes. Variable mutation rates are modeled as lognormally distributed with mean1.2×10^{-7}, as described in Harpak et al. 2016.

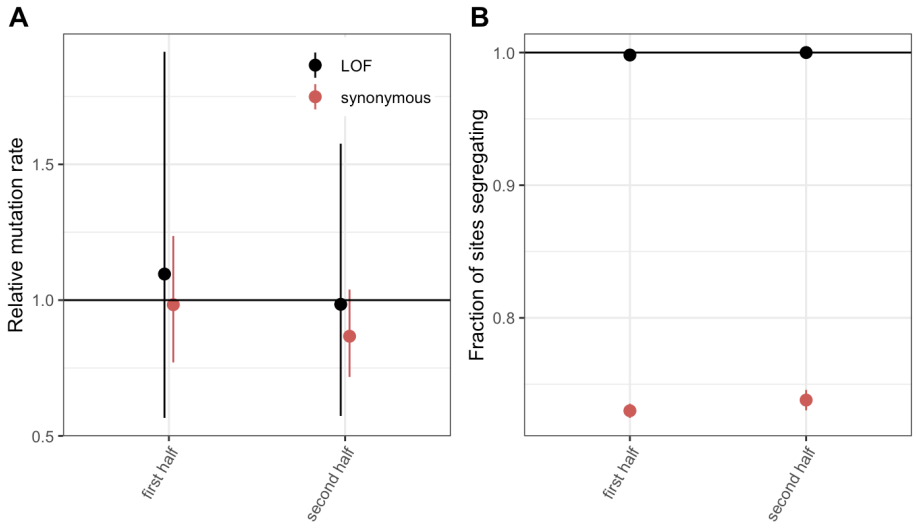

Figure 2.10: (a) DNM rates for synonymous and LOF CpG transitions at highly methylated sites in exons that constitute the first vs. second halves of canonical protein coding transcripts, normalized by the total DNM rate in exons, with 95% Poisson confidence intervals. (b) Fraction of highly methylated CpG sites that are segregating as a synonymous or LOF C/T polymorphism in exons that constitute the first vs. second halves of canonical protein coding transcripts, relative to the fraction of all synonymous sites segregating. Error bars are 95% confidence intervals assuming the number of segregating sites is binomially distributed (see Methods). LOF variants are defined as stop-gained and splice donor/acceptor variants that do not fall near the end of the transcript, and meet the other criteria to be classified as "high-confidence" loss-of-function in gnomAD.

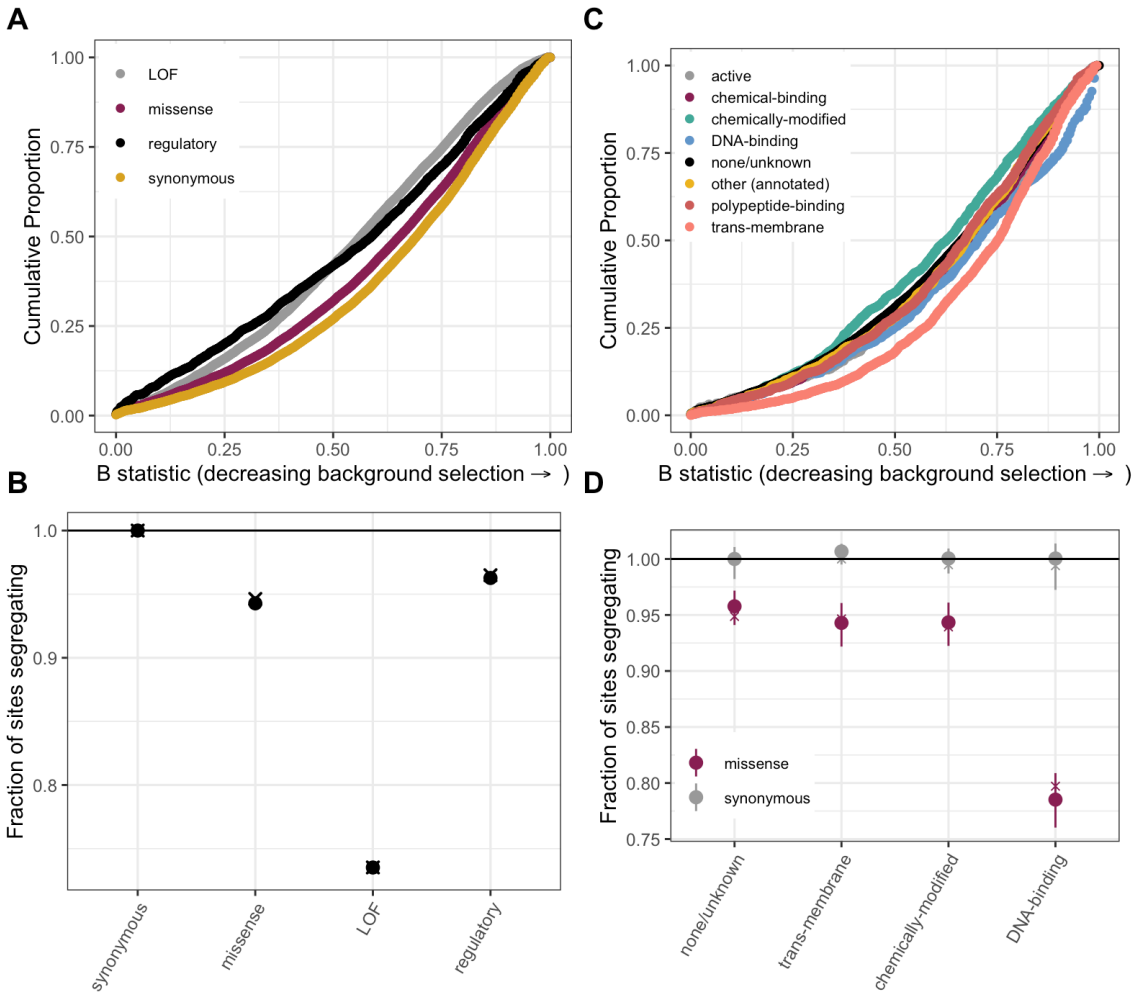

Figure 2.11: (a) Cumulative distribution of the B-statistic from McVicker et al., 2009 for all possible CpG transitions at highly methylated sites by annotation class. (b) Fraction of highly methylated CpG sites that are segregating as a C/T polymorphism in an annotation class, relative to the fraction of synonymous sites segregating, after matching the distribution of the B-statistic across annotations. The fraction segregating without matching for B-statistics is denoted by crosses, to enable comparison to **Figure 2.3**. Regulatory variants include non-LOF splice region variants and UTRs. (c) Cumulative distribution of the B-statistic for all possible CpG transitions at highly methylated sites by functional class. (d) The proportion of synonymous and missense segregating C/T polymorphisms for four functional classes, after matching the distribution of the B-statistic across categories. Error bars are 95% confidence intervals assuming the number of segregating sites is binomially distributed. The fraction segregating without matching for B-statistics is denoted by crosses, to enable comparison to **Figure 2.3**.

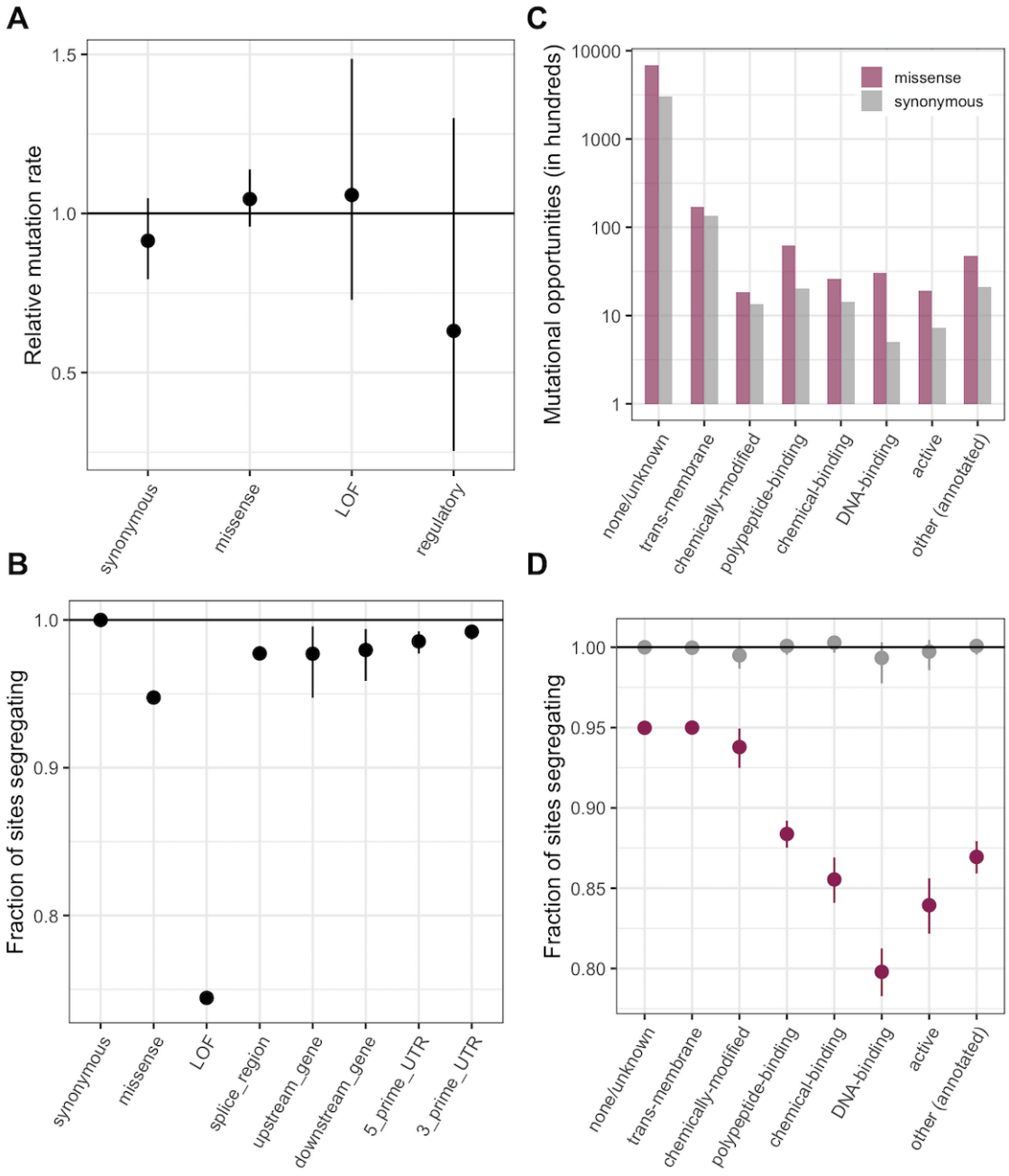

Figure 2.12: The analyses in **Figure 2.3**, but with annotations obtained using the worst consequence in protein coding transcripts by severity, instead of canonical transcripts; the order of preference by which functional sites are assigned to a single category is detailed in Methods.

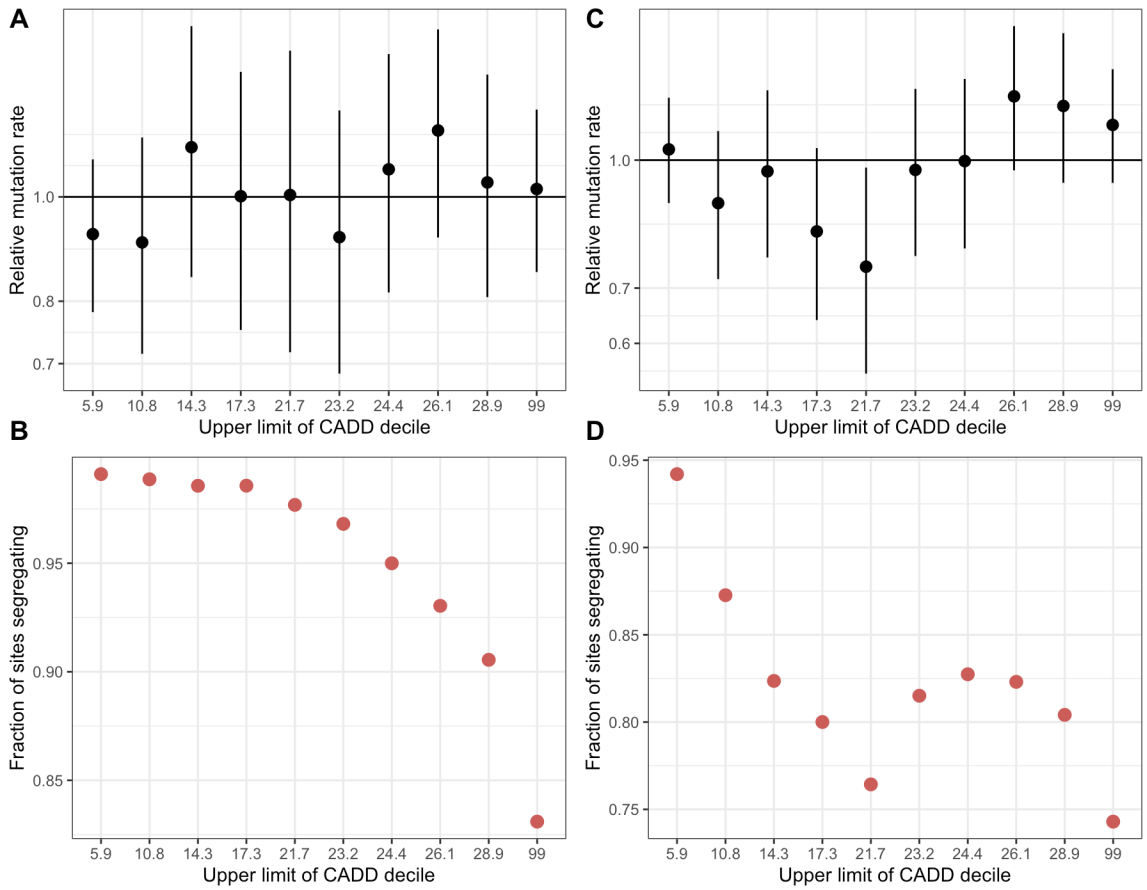

Figure 2.13: (a) De novo C>T mutation rate at highly methylated CpGs in deciles of CADD scores in exons, normalized by the total rate of highly methylated CpG transitions in exons. Error bars reflect the 95% Poisson confidence interval around mutation counts in each group. (b) Fraction of highly methylated CpG sites that are segregating as a C/T polymorphism in a CADD score decile, relative to the fraction of synonymous sites segregating. (c) The same as (a) but for C>T mutations at all CpG sites, including unmethylated and less methylated CpGs as well as highly methylated ones. (d) The same as (b) but for C>T mutations at all CpG sites.

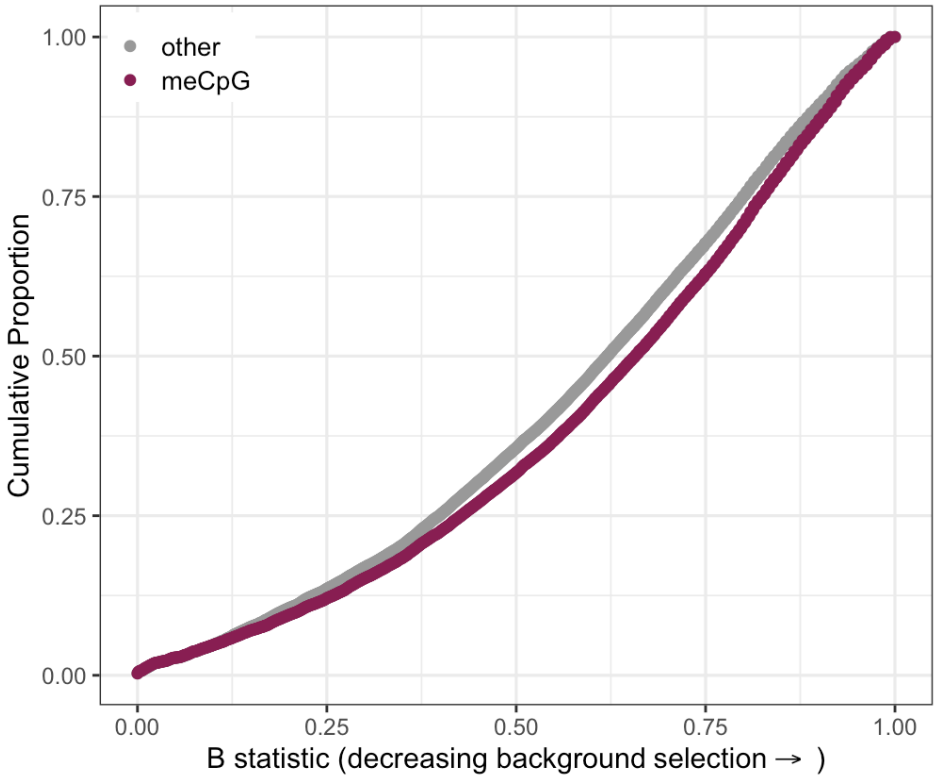

Figure 2.14: Cumulative distribution of the B-statistic from McVicker et al., 2009 for highly methylated CpG sites vs. all other types of sites in exons (ks test p-value $< 10^{-5}$).

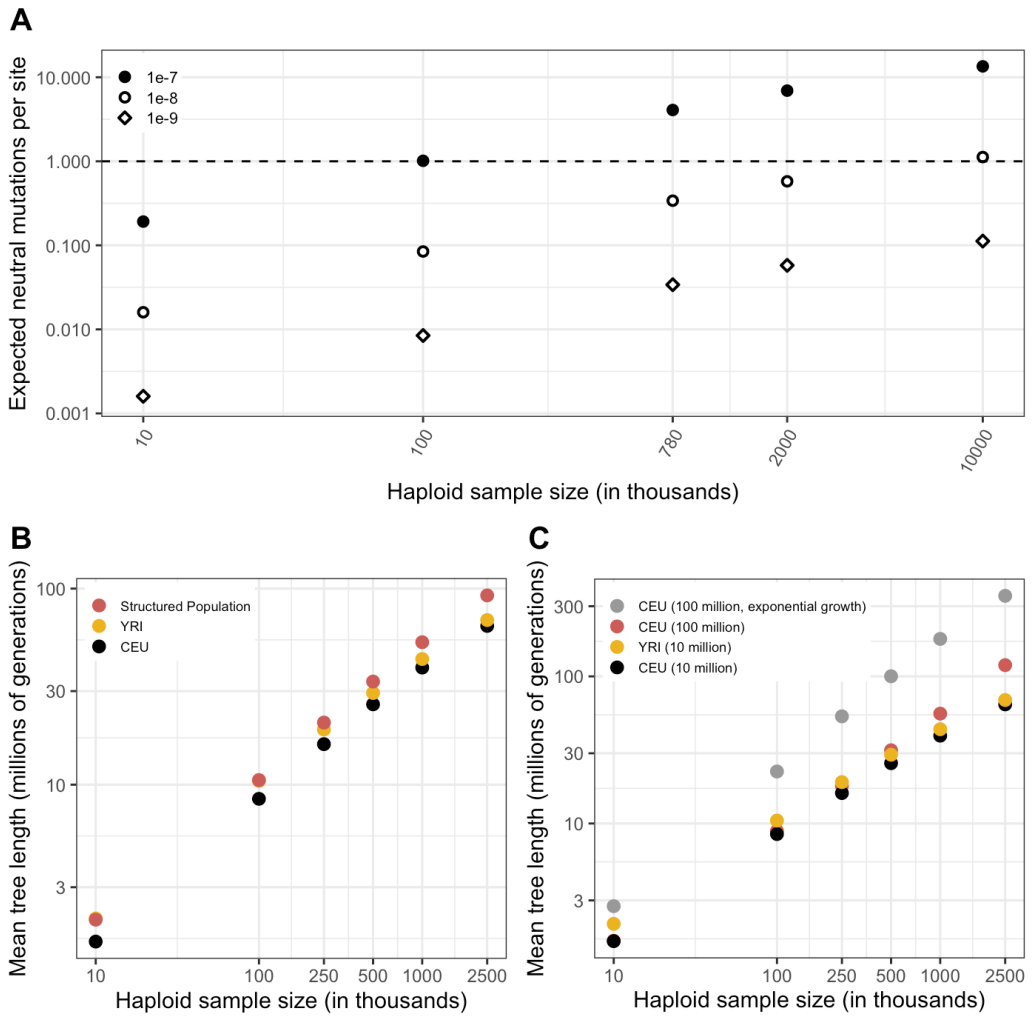

Figure 2.15: (a) The expected number of neutral mutations in a sample, for three mutation rates, calculated as the expected length of the genealogy (averaged over 20 simulations) for a CEU sample $\times$mutation rate. (b) A comparison of mean tree lengths for four variations on the Schiffels-Durbin demographic models for CEU and YRI populations, namely, YRI demographic history with a recent N_e of 10 million for the last 50 generations, CEU demographic history for 50,000 generations with a recent N_e of 10 million or 100 million, and CEU demographic history with 5% exponential growth for the past 200 generations. (b) A comparison of mean tree lengths for samples from YRI and CEU populations, and samples from a structured population derived from an ancestral population 2,000 generations ago.

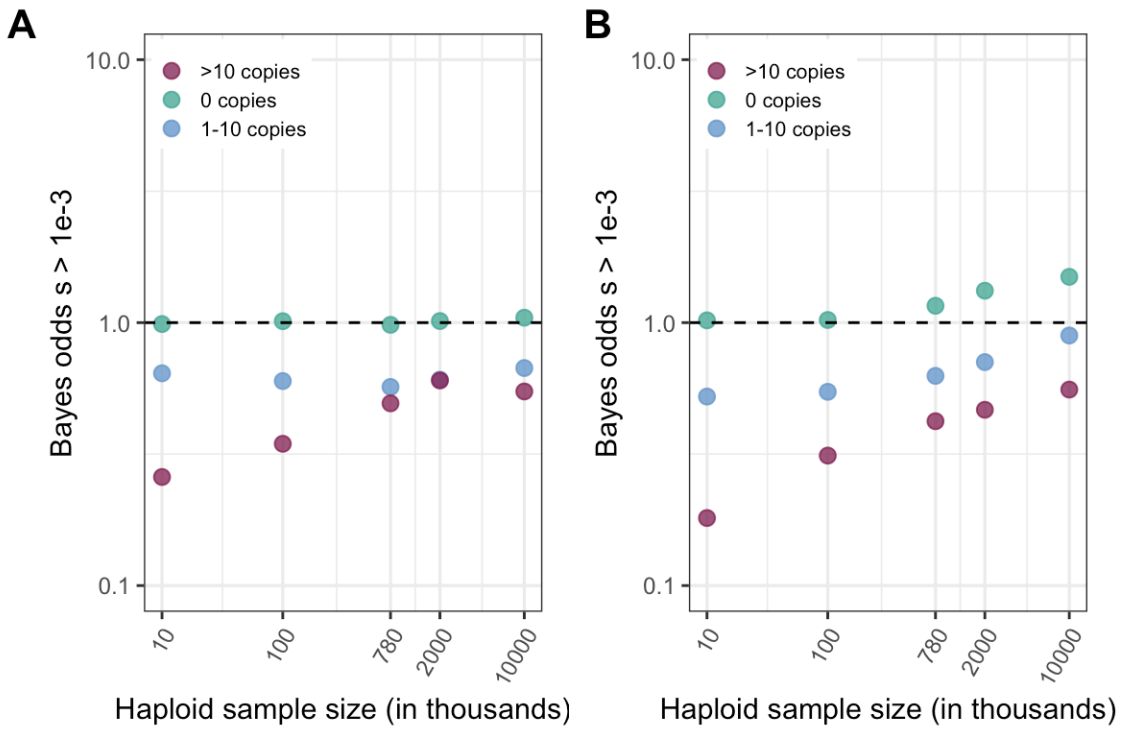

Figure 2.16: For various sample sizes, the Bayes odds of s > 0.001 (h=0.5) for a mutation observed at 0, 1-10, or >10 copies, where the prior distribution of s is log-uniform over [10^{-7},1]. The odds are calculated from 15,000 draws from the prior and posterior distributions (a) At a site with mutation rate ~ 10^{-9} (b) At a site with mutation rate ~ 10^{-8}.

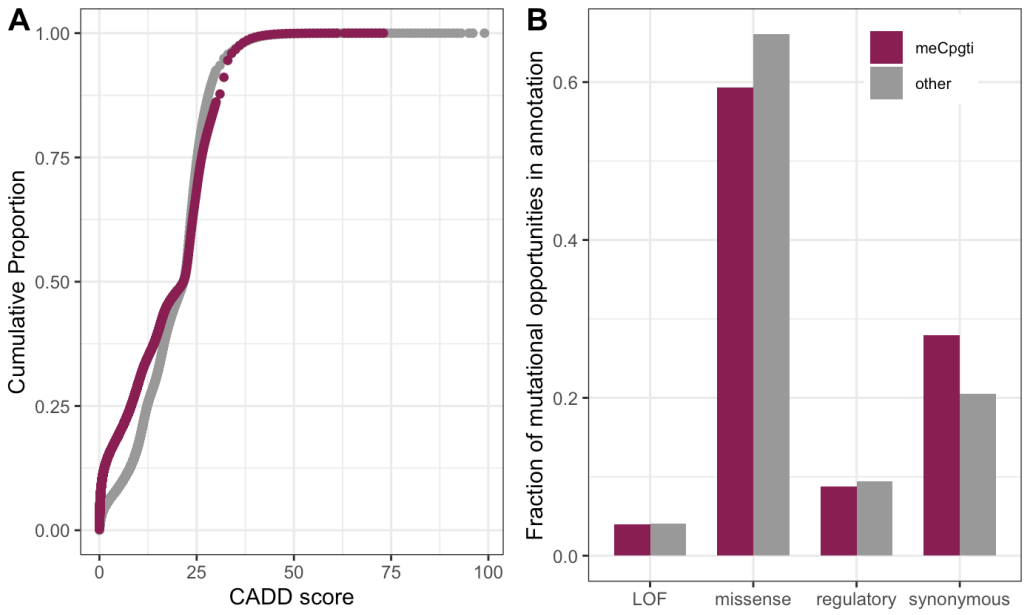

Figure 2.17: (a) Distribution of CADD scores for 1.1 million mutational opportunities for CpG transitions vs. 90 million other mutational opportunities in exons (p-value from a Kolmogorov-Smirnov test $<< 10^{-5}$). (b) Fraction of mutational opportunities for CpG transitions vs. all other mutational opportunities in exons by their putative functional effect. The difference is statistically significant for missense, regulatory, and synonymous categories (Fisher exact test p-value $<< 10^{-5}$) but not for the LOF class (p-value=0.06).

Table 2.1: Sources of annotation data for exons.

Annotation type	**Source**
Exon coordinates	http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz
Exon annotations	VEP v87 using Gencode v19 Ranks:https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html
WGS covered regions and exome target regions (gnomAD v2.1.1)	https://gnomad.broadinstitute.org/downloads
Exome target regions (UK Biobank)	https://biobank.ndph.ox.ac.uk/ukb/ukb/auxdata/xgen_plus_spikein.GRCh38.bed (liftovered to hg19)
CpG methylation Testis (sperm)	GEO Accession GSM1127119 (https://www.ncbi.nlm.nih.gov/geo/)
CpG methylation Ovary	GEO Accession GSM1010980 (https://www.ncbi.nlm.nih.gov/geo/)
CADD	CADD v1.4 (https://cadd.gs.washington.edu/download)
Functional site annotations	http://www.prot2hg.com/
De novo mutations	https://doi.org/10.1126/science.aau1043 (Data S5)
Polymorphism data	gnomAD: https://gnomad.broadinstitute.org/downloads; UK Biobank: https://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=23155; DiscovEHR: http://www.discovehrshare.com/downloads; 1000 Genomes:http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Future directions

The availability of hundreds of thousands of human whole exome and whole genome sequences has made it possible to observe human genetic variation on an unprecedented scale, and in many cases, to infer evolutionary parameters of interest [9, 10, 11]. The distribution of fitness effects at sites in the genome, for instance, has long been of great interest to evolutionary biologists as well as human geneticists [7]. Current samples are already sufficiently large to directly infer the strength of selection from patterns of observed genetic variation for loss-of-function (LOF) variants in genes: these estimates have proven useful in triaging genetic causes of undiagnosed severe childhood diseases [123, 136]. LOF variants are a special case where each gene can be treated as a single locus at which many possible LOF mutations have the same fitness impact: the rate at which average entire gene experiences loss-of-function events is $\sim 10^{-6}$ per generation, large relative to the genealogical history of current samples, and almost ten-fold greater than the most mutable sites in the genome. As discussed in Chapter 2, the distribution of fitness effects at ~1 million highly methylated CpG sites in the exome, and by proxy, at other sites in exons, should be within reach with samples of a few million individuals. A more precise characterization of mutation rate

variation within the class of methylated CpGs (and a reliable demographic model) will be needed to infer the strength of selection at these sites, however.

With regard to characterizing the sources of mutation for CpGs as well as other types of sites, identifying germline mutational signatures associated with particular damage and repair mechanisms promises to yield valuable insights. In addition to distributions of biochemical markers in the genome, some examples of which are highlighted in Chapter 1, patterns of strand-asymmetry in genetic variation can be used to identify signatures of damage and repair associated with transcription and replication [137]. As ever-increasing amounts of functional information from human cell lines, tumor samples, and model organisms (e.g., [138]) becomes available, statistical approaches that systematically identify mutational signatures in polymorphism data [139] and interpret them in the light of this functional information will be immensely valuable in elucidating the kinds of mutational processes active in the germline and their contributions to mutation rate variation for different mutation types.

Moreover, as the number of parent-offspring trios sequenced increases, it should become possible to estimate de novo mutation rates for an increasing, though still limited, number of broad mutational contexts – as we showed in Chapter 2, this is already extremely useful for different subsets of CpG sites. More importantly, inter-individual variation in de novo mutations and the dependence of different types of mutations on sex and age is a promising source of insights into mutational mechanisms [49]. Because de novo mutations reflect one generation of germline development, learning more about the biological processes of germline development – for instance, the number of cell divisions at different stages of development, and the trajectory of

methylation and time series information on other biochemical features of germ cells – and how they differ in males and females with age will help inform our understanding of how different types of mutations arise during development. Finally, at least some variation in polymorphism levels at different sites in the genome is attributable to mutation rates having varied over evolutionary time, and even on relatively short time scales [110, 52, 111, 112]. Understanding how much this has contributed to the distribution of diversity for different mutation types will be of relevance in interpreting signatures of mutational processes in polymorphism data [140]. In summary, both characterizing the sources of mutation and inferring fitness effects in the genome will benefit from development of models that capture mutation rate variation over the genome and over different time scales, as well as from increasing amounts of functional and genetic variation data.

Bibliography

[1] J. F. Crow and M. Kimura. *An Introduction to Population Genetics: Theory and Applications*. Ed. by Motoo Kimura 1924-. New York: Harper & Row, 1970, p. 591.

[2] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, 1983.

[3] Martin Kreitman. "The neutral theory is dead. Long live the neutral theory." In: *BioEssays* 18.8 (Aug. 1996), pp. 678–683.

[4] Sarah P Otto. "Detecting the form of selection from DNA sequence data." In: *Trends in Genetics* 16.12 (Dec. 2000), pp. 526–529.

[5] R.R. Hudson. "Gene genealogies and the coalescent process." In: *Oxford surveys in evolutionary biology* 7.1 (1990), p. 44.

[6] Brian Charlesworth and Deborah Charlesworth. *Elements of evolutionary genetics. Roberts and Company Publishers*. 2010.

[7] Adam Eyre-Walker and Peter D Keightley. "The distribution of fitness effects of new mutations." In: *Nature Reviews Genetics* 8.8 (2007), pp. 610–618.

[8] J. Felsenstein. "PHYLIP - Phylogeny Inference Package (Version 3.2)." In: *Cladistics* 5 (1989), pp. 164–166.

[9] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences." In: *Nature Genetics* 46.8 (2014), pp. 919–925.

[10] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples." In: *BioRxiv* (2019), p. 550558.

[11] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets." In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[12] Dario Boffelli et al. “Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome.” In: *Science* 299.5611 (Feb. 2003), 1391 LP –1394.

[13] Konrad J Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans.” In: *Nature* 581.7809 (2020), pp. 434–443.

[14] Monkol Lek et al. “Analysis of protein-coding genetic variation in 60,706 humans.” In: *Nature* 536.7616 (2016), pp. 285–291.

[15] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.” eng. In: *Genome research* 15.8 (Aug. 2005), pp. 1034–1050.

[16] Katherine S Pollard et al. “Detection of nonneutral substitution rates on mammalian phylogenies.” eng. In: *Genome research* 20.1 (Jan. 2010), pp. 110–121.

[17] Xin Yi et al. “Sequencing of 50 human exomes reveals adaptation to high altitude.” eng. In: *Science (New York, N.Y.)* 329.5987 (July 2010), pp. 75–78.

[18] J B S Haldane. “The rate of spontaneous mutation of a human gene.” In: *Journal of Genetics* 83.3 (2004), pp. 235–244.

[19] Robert H Waterson et al. “Initial sequence of the chimpanzee genome and comparison with the human genome.” In: *Nature* 437.7055 (2005), pp. 69–87.

[20] Michael W. Nachman and Susan L. Crowell. “Estimate of the Mutation Rate per Nucleotide in Humans.” In: *Genetics* 156.1 (2000), pp. 297–304.

[21] Augustine Kong et al. “Rate of de novo mutations and the importance of father’s age to disease risk.” In: *Nature* 488.7412 (2012), pp. 471–475.

[22] Hákon Jónsson et al. “Parental influence on human germline de novo mutations in 1,548 trios from Iceland.” In: *Nature* 549.7673 (Sept. 2017), pp. 519–522.

[23] Bruce K Duncan and Jeffrey H Miller. “Mutagenic deamination of cytosine residues in DNA.” In: *Nature* 287.5782 (1980), pp. 560–561.

[24] A Harpak, A Bhaskar, and J K Pritchard. “Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans.” In: *PLoS Genetics* 12.12 (2016).

[25] Alan Hodgkinson, Emmanuel Ladoukakis, and Adam Eyre-Walker. “Cryptic Variation in the Human Mutation Rate.” In: *PLOS Biology* 7.2 (Feb. 2009), pp. 1–7.

[26] Bjarni V Halldorsson et al. “Characterizing mutagenic effects of recombination through a sequence-level genetic map.” In: *Science* 363.6425 (Jan. 2019), eaau1043.

[27] Ichiro Hiratani et al. “Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis.” In: *Genome Research* 20.2 (Feb. 2010), pp. 155–169.

[28] Amnon Koren et al. “Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation.” In: *The American Journal of Human Genetics* 91.6 (Dec. 2012), pp. 1033–1040.

[29] Florencia Pratto et al. “Recombination initiation maps of individual human genomes.” In: *Science* 346.6211 (Nov. 2014).

[30] Bradley E Bernstein et al. “The NIH Roadmap Epigenomics Mapping Consortium.” eng. In: *Nature biotechnology* 28.10 (Oct. 2010), pp. 1045–1048.

[31] Michael M. Hoffman et al. “Integrative annotation of chromatin elements from ENCODE data.” In: *Nucleic Acids Research* 41.2 (Dec. 2012), pp. 827–841.

[32] Francis Blokzijl et al. “Tissue-specific mutation accumulation in human adult stem cells during life.” In: *Nature* 538 (Oct. 2016), p. 260.

[33] Christopher Greenman et al. “Patterns of somatic mutation in human cancer genomes.” In: *Nature* 446 (Mar. 2007), p. 153.

[34] Alan Hodgkinson, Ying Chen, and Adam Eyre-Walker. “The large-scale distribution of somatic mutations in cancer genomes.” In: *Human Mutation* 33.1 (Sept. 2011), pp. 136–143.

[35] Lin Liu, Subhajyoti De, and Franziska Michor. “DNA replication timing and higher-order nuclear organization determine single nucleotide substitution patterns in cancer genomes.” In: *Nature communications* 4 (2013), p. 1502.

[36] Erin D Pleasance et al. “A comprehensive catalogue of somatic mutations from a human cancer genome.” In: *Nature* 463 (Dec. 2009), p. 191.

[37] Paz Polak et al. “Cell-of-origin chromatin organization shapes the mutational landscape of cancer.” In: *Nature* 518 (Feb. 2015), p. 360.

[38] Alan F Rubin and Phil Green. “Mutation patterns in cancer genomes.” In: *Proceedings of the National Academy of Sciences* 106.51 (Dec. 2009), 21766 LP –21770.

[39] Benjamin Schuster-Böckler and Ben Lehner. "Chromatin organization is a major influence on regional mutation rates in human cancer cells." In: *Nature* 488 (July 2012), p. 504.

[40] Fran Supek and Ben Lehner. "Differential DNA mismatch repair underlies mutation rate variation across the human genome." In: *Nature* 521.7550 (May 2015), pp. 81–84.

[41] Yong H Woo and Wen-Hsiung Li. "DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes." In: *Nature Communications* 3 (Aug. 2012), p. 1004.

[42] Ludmil Alexandrov et al. "The Repertoire of Mutational Signatures in Human Cancer." In: *bioRxiv* (Jan. 2018), p. 322859.

[43] Ludmil B. Alexandrov et al. "Signatures of mutational processes in human cancer." In: *Nature* 500.7463 (2013), pp. 415–421.

[44] Serena Nik-Zainal et al. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." In: *Nature* 534 (May 2016), p. 47.

[45] Erin D Pleasance et al. "A small-cell lung cancer genome with complex signatures of tobacco exposure." In: *Nature* 463 (Dec. 2009), p. 184.

[46] Jedidiah Carlson et al. "Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans." In: *Nature Communications* 9.1 (2018), p. 3753.

[47] Raheleh Rahbari et al. "Timing, rates and spectra of human germline mutation." In: *Nature Genetics* 48.December (2015), pp. 1–11.

[48] Valerie M. Schaibley et al. "The influence of genomic context on mutation patterns in the human genome inferred from rare variants." In: *Genome Research* (2013).

[49] Ziyue Gao et al. "Overlooked roles of DNA damage and maternal age in generating human germline mutations." In: *Proceedings of the National Academy of Sciences* 116.19 (May 2019), 9491 LP –9500.

[50] Laure Ségurel, Minyoung J. Wyman, and Molly Przeworski. "Determinants of Mutation Rate Variation in the Human Germline." In: *Annual Review of Genomics and Human Genetics* 15.1 (2014), pp. 47–70.

[51] Michael Lynch et al. "Genetic drift, selection and the evolution of the mutation rate." In: *Nature Reviews Genetics* 17 (Oct. 2016), p. 704.

[52] Kelley Harris and Jonathan K. Pritchard. “Rapid evolution of the human mutation spectrum.” In: *eLife* 6 (Apr. 2017).

[53] Michael B Burns, Nuri A Temiz, and Reuben S Harris. “Evidence for APOBEC3B mutagenesis in multiple human cancers.” In: *Nature Genetics* 45 (July 2013), p. 977.

[54] Serena Nik-Zainal et al. “Mutational Processes Molding the Genomes of 21 Breast Cancers.” In: *Cell* 149.5 (May 2012), pp. 979–993.

[55] Steven A Roberts et al. “An APOBEC Cytidine Deaminase Mutagenesis Pattern is Widespread in Human Cancers.” In: *Nature genetics* 45.9 (Sept. 2013), pp. 970–976.

[56] James F Crow. “The origins, patterns and implications of human spontaneous mutation.” In: *Nature Reviews Genetics* 1 (Oct. 2000), p. 40.

[57] Adetunji P Fayomi and Kyle E Orwig. “Spermatogonial stem cells and spermatogenesis in mice, monkeys and men.” In: *Stem Cell Research* 29 (2018), pp. 207–214.

[58] Meisha A Morelli and Paula E Cohen. “Not all germ cells are created equal: Aspects of sexual dimorphism in mammalian meiosis.” English. In: *Reproduction* 130.6 (2005), pp. 761–781.

[59] Walfred W C Tang et al. “Specification and epigenetic programming of the human germ line.” In: *Nature Reviews Genetics* 17 (Aug. 2016), p. 585.

[60] Wolf Reik, Wendy Dean, and Jörn Walter. “Epigenetic Reprogramming in Mammalian Development.” In: *Science* 293.5532 (Aug. 2001), 1089 LP –1093.

[61] Tegan B Smith et al. “The presence of a truncated base excision repair pathway in human spermatozoa that is mediated by OGG1.” In: *Journal of Cell Science* 126.6 (Mar. 2013), 1488 LP –1497.

[62] Jack N Fenner. “Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies.” In: *American Journal of Physical Anthropology* 128.2 (Mar. 2005), pp. 415–423.

[63] Laurent C Francioli et al. “Genome-wide patterns and properties of de novo mutations in humans.” In: *Nature Genetics* 47 (May 2015), p. 822.

[64] Jakob M Goldmann et al. “Parent-of-origin-specific signatures of de novo mutations.” In: *Nature Genetics* 48.8 (2016), pp. 935–939.

[65] Thomas C A Smith, Peter F Arndt, and Adam Eyre-Walker. "Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans." In: *PLOS Genetics* 14.3 (Mar. 2018), e1007254.

[66] Chen Chen et al. "Contrasting Determinants of Mutation Rates in Germline and Soma." In: *Genetics* (July 2017).

[67] Alan Hodgkinson and Adam Eyre-Walker. "Variation in the mutation rate across mammalian genomes." In: *Nature Reviews Genetics* 12 (Oct. 2011), p. 756.

[68] Kateryna D Makova and Wen-Hsiung Li. "Strong male-driven evolution of DNA sequences in humans and apes." In: *Nature* 416 (Apr. 2002), p. 624.

[69] Lawrence C Shimmin, Benny Hung-Junn Chang, and Wen-Hsiung Li. "Male-driven evolution of DNA sequences." In: *Nature* 362 (Apr. 1993), p. 745.

[70] Konrad J Karczewski et al. "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes." In: *bioRxiv* 581.7809 (Jan. 2019), pp. 434–443.

[71] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (2015), pp. 68–74.

[72] John A Stamatoyannopoulos et al. "Human mutation rate associated with DNA replication timing." In: *Nature Genetics* 41 (Mar. 2009), p. 393.

[73] Tyrone Ryba et al. "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types." In: *Genome Research* 20.6 (June 2010), pp. 761–770.

[74] Aimée M Deaton and Adrian Bird. "CpG islands and the regulation of transcription." In: *Genes & Development* 25.10 (May 2011), pp. 1010–1022.

[75] Hao Wu et al. "Redefining CpG islands using hidden Markov models." In: *Biostatistics (Oxford, England)* 11.3 (July 2010), pp. 499–514.

[76] Sheila S David, Valerie L O'Shea, and Sucharita Kundu. "Base-excision repair of oxidative DNA damage." In: *Nature* 447 (June 2007), p. 941.

[77] William L Neeley and John M Essigmann. "Mechanisms of Formation, Genotoxicity, and Mutation of Guanine Oxidation Products." In: *Chemical Research in Toxicology* 19.4 (Apr. 2006), pp. 491–505.

[78] Geoffry N De Iuliis et al. “DNA Damage in Human Spermatozoa Is Highly Correlated with the Efficiency of Chromatin Remodeling and the Formation of 8-Hydroxy-2 -Deoxyguanosine, a Marker of Oxidative Stress1.” In: *Biology of Reproduction* 81.3 (Sept. 2009), pp. 517–524.

[79] Chad Harland et al. “Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle.” In: *bioRxiv* (Jan. 2017), p. 79863.

[80] August Y Huang et al. “Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals.” In: *Cell Research* 24 (Oct. 2014), p. 1311.

[81] Young Seok Ju et al. “Somatic mutations reveal asymmetric cellular dynamics in the early human embryo.” In: *Nature* 543 (Mar. 2017), p. 714.

[82] Wolf Reik and Anne C Ferguson-Smith. “The X-inactivation yo-yo.” In: *Nature* 438 (Nov. 2005), p. 297.

[83] Amnon Koren et al. “Genetic variation in human DNA replication timing.” In: *Cell* 159.5 (Nov. 2014), pp. 1015–1026.

[84] Edith Heard and Christine M Disteche. “Dosage compensation in mammals: fine-tuning the expression of the X chromosome.” In: *Genes & Development* 20.14 (July 2006), pp. 1848–1867.

[85] Di Kim Nguyen and Christine M Disteche. “Dosage compensation of the active X chromosome in mammals.” In: *Nature Genetics* 38 (Dec. 2005), p. 47.

[86] Bernhard Payer, Jeannie T Lee, and Satoshi H Namekawa. “X-inactivation and X-reactivation: epigenetic hallmarks of mammalian reproduction and pluripotent stem cells.” In: *Human genetics* 130.2 (Aug. 2011), pp. 265–280.

[87] Mahesh N Sangrithi and James M A Turner. “Mammalian X Chromosome Dosage Compensation: Perspectives From the Germ Line.” In: *BioEssays* 40.6 (May 2018), p. 1800024.

[88] Laura Carrel and Huntington F Willard. “X-inactivation profile reveals extensive variability in X-linked gene expression in females.” In: *Nature* 434 (Mar. 2005), p. 400.

[89] Taru Tukiainen et al. “Landscape of X chromosome inactivation across human tissues.” In: *Nature* 550 (Oct. 2017), p. 244.

[90] C Allegrucci and L E Young. “Differences between human embryonic stem cell lines.” In: *Human Reproduction Update* 13.2 (Mar. 2007), pp. 103–120.

[91] Céline Vallot et al. “Erosion of X Chromosome Inactivation in Human Pluripotent Cells Initiates with XACT Coating and Depends on a Specific Heterochromatin Landscape.” In: *Cell Stem Cell* 16.5 (2015), pp. 533–546.

[92] Sanjeet Patel et al. “Human embryonic stem cells do not change their X-inactivation status during differentiation.” In: *Cell reports* 18.1 (Jan. 2017), pp. 54–67.

[93] Guy Amster et al. “Changes in life history and population size can explain the relative neutral diversity levels on X and autosomes in extant human populations.” In: *Proceedings of the National Academy of Sciences* 117 (Aug. 2020), p. 201915664.

[94] Boubou Diagouraga et al. “PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites.” In: *Molecular Cell* 69.5 (Mar. 2018), 853–865.e6.

[95] Simon Myers et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination.” In: *Science* 327.5967 (Feb. 2010), pp. 876–879.

[96] Frédéric Baudat, Yukiko Imai, and Bernard de Massy. “Meiotic recombination in mammals: localization and regulation.” In: *Nature Reviews Genetics* 14 (Oct. 2013), p. 794.

[97] Ran Li et al. “A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination.” In: *bioRxiv* (Jan. 2018).

[98] Liisa Kauppi et al. “Distinct Properties of the XY Pseudoautosomal Region Crucial for Male Meiosis.” In: *Science* 331.6019 (Feb. 2011), 916 LP –920.

[99] Lin-Yu Lu and Xiaochun Yu. “Double-strand break repair on sex chromosomes: challenges during male meiotic prophase.” In: *Cell Cycle* 14.4 (Jan. 2015), pp. 516–525.

[100] Peter B Moens et al. “Rad51 immunocytology in rat and mouse spermatocytes and oocytes.” In: *Chromosoma* 106.4 (1997), pp. 207–215.

[101] Anjali G Hinch et al. “Recombination in the Human Pseudoautosomal Region PAR1.” In: *PLOS Genetics* 10.7 (July 2014), e1004503.

[102] Liisa Kauppi, Maria Jasin, and Scott Keeney. “The tricky path to recombining X and Y chromosomes in meiosis.” In: *Annals of the New York Academy of Sciences* 1267 (Sept. 2012), pp. 18–23.

[103] Julian Lange et al. “The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair.” In: *Cell* 167.3 (Oct. 2016), 695–708.e16.

[104] A Helena Mangs and Brian J Morris. “The Human Pseudoautosomal Region (PAR): Origin, Function and Future.” In: *Current Genomics* 8.2 (Apr. 2007), pp. 129–136.

[105] Claude Bhérer, Christopher L Campbell, and Adam Auton. “Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales.” eng. In: *Nature communications* 8 (Apr. 2017), p. 14994.

[106] Irene da Cruz et al. “Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage.” In: *BMC Genomics* 17.1 (2016), p. 294.

[107] Andrea Enguita-Marruedo et al. “Transition from a meiotic to a somatic-like DNA damage response during the pachytene stage in mouse meiosis.” In: *PLOS Genetics* 15.1 (Jan. 2019), e1007439.

[108] Gennady Margolin et al. “Integrated transcriptome analysis of mouse spermatogenesis.” In: *BMC Genomics* 15.1 (2014), p. 39.

[109] Vijayalakshmi V Subramanian et al. “Chromosome Synapsis Alleviates Mek1-Dependent Suppression of Meiotic DNA Repair.” In: *PLOS Biology* 14.2 (Feb. 2016), e1002369.

[110] Kelley Harris. “Evidence for recent, population-specific evolution of the human mutation rate.” In: *Proceedings of the National Academy of Sciences* 112.11 (2015), pp. 3439–3444.

[111] Iain Mathieson and David Reich. “Differences in the rare variant spectrum among human populations.” In: *PLOS Genetics* 13.2 (Feb. 2017), e1006581.

[112] Vagheesh M. Narasimhan et al. “Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes.” In: *Nature Communications* 8.1 (2017), p. 303.

[113] Nicola Barban et al. “Genome-wide analysis identifies 12 loci influencing human reproductive behavior.” In: *Nature Genetics* (2016).

[114] John R B Perry et al. “Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche.” In: *Nature* 514 (July 2014), p. 92.

[115] James Taylor et al. “Strong and Weak Male Mutation Bias at Different Sites in the Primate Genomes: Insights from the Human-Chimpanzee Comparison.” In: *Molecular Biology and Evolution* 23.3 (Mar. 2006), pp. 565–573.

[116] Adam Siepel and David Haussler. “Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood.” In: *Molecular Biology and Evolution* 21.3 (Mar. 2004), pp. 468–488.

[117] Swapan Mallick et al. “The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.” In: *Nature* 538.7624 (2016), pp. 201–206.

[118] Augustine Kong et al. “Fine-scale recombination rate differences between sexes, populations and individuals.” In: *Nature* 467 (Oct. 2010), p. 1099.

[119] Frederick E Dewey et al. “Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study.” In: *Science* 354.6319 (Dec. 2016), aaf6814.

[120] Joseph D Szustakowski et al. “Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank.” In: *medRxiv* (Jan. 2020), p. 2020.11.02.20222232.

[121] Cristopher V Van Hout et al. “Exome sequencing and characterization of 49,960 individuals in the UK Biobank.” In: *Nature* 586.7831 (2020), pp. 749–756.

[122] Christopher A Cassa et al. “Estimating the selective effects of heterozygous protein-truncating variants from human exome data.” In: *Nature Genetics* 49.5 (2017), pp. 806–810.

[123] Joanna Kaplanis et al. “Evidence for 28 genetic disorders discovered by combining healthcare and research data.” In: *Nature* 586.7831 (2020), pp. 757–762.

[124] Zachary L Fuller et al. “Measuring intolerance to mutation in human genetics.” In: *Nature Genetics* 51.5 (2019), pp. 772–776.

[125] Daniel Taliun et al. “Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.” In: *Nature* 590.7845 (2021), pp. 290–299.

[126] Graham McVicker et al. “Widespread Genomic Signatures of Natural Selection in Hominid Evolution.” In: *PLOS Genetics* 5.5 (May 2009), e1000471.

[127] Philipp Rentzsch et al. “CADD: predicting the deleteriousness of variants throughout the human genome.” In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D886–D894.

[128] Anand Bhaskar, Andrew G Clark, and Yun S Song. “Distortion of genealogical properties when the sample is very large.” In: *Proceedings of the National Academy of Sciences* 111.6 (Feb. 2014), 2385 LP –2390.

[129] Adam R Boyko et al. “Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome.” In: *PLOS Genetics* 4.5 (May 2008), e1000083.

[130] Adam Eyre-Walker, Megan Woolfit, and Ted Phelps. “The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans.” In: *Genetics* 173.2 (June 2006), 891 LP –900.

[131] Fernando Racimo and Joshua G Schraiber. “Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms.” In: *PLOS Genetics* 10.11 (Nov. 2014), e1004697.

[132] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data.” In: *Nature* 562.7726 (2018), p. 203.

[133] David Stanek et al. “Prot2HG: a database of protein domains mapped to the human genome.” In: *Database* 2020 (Jan. 2020).

[134] Yuval B Simons et al. “The deleterious mutation load is insensitive to recent population history.” eng. In: *Nature genetics* 46.3 (Mar. 2014), pp. 220–224.

[135] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. “Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes.” In: *PLOS Computational Biology* 12.5 (May 2016), e1004842.

[136] Donate Weghorn et al. “Applicability of the Mutation–Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans.” In: *Molecular Biology and Evolution* 36.8 (Apr. 2019), pp. 1701–1710.

[137] Vladimir B Seplyarskiy et al. “Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations.” In: *Nature Genetics* 51.1 (2019), pp. 36–41.

[138] Nadezda V Volkova et al. “Mutational signatures are jointly shaped by DNA damage and repair.” In: *Nature Communications* 11.1 (2020), p. 2169.

[139] Vladimir B Seplyarskiy et al. "Population sequencing data reveal a compendium of mutational processes in human germline." In: *bioRxiv* (Jan. 2020), p. 2020.01.10.893024.

[140] Jedidiah Carlson, William S DeWitt, and Kelley Harris. "Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation." In: *Current Opinion in Genetics & Development* 62 (2020), pp. 50–57.